一 本 最 接 地 气 的 “ 大 数 据 ” 类 图 书

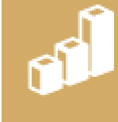


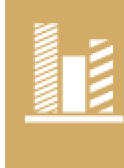


大数据

从海量到精准

李 军◎编著





- 12大行业领域应用
- 15章大数据专题精讲
- 110多个经典专家提醒
- 120个大数据应用案例
- 150多张图片全程图解

帮助读者在最短的时间内掌控大数据的秘密

这本书让读者了解了什么是大数据，大数据的潜在商业价值，大数据无处不在的应用，大数据对人类生活带来哪些影响，大数据与个人隐私及公共安全等。同时，这本书对于公共政策、信息科学、社会科学等领域的交叉融合也具有启发意义。

清华大学出版社





大数据

从海量到精准

李　军◎编著

清华大学出版社
北　京

内 容 简 介

本书共分为15章，具体内容包括入门：大数据的基本概念；价值：大数据商业变革；架构：大数据基础设施；掌握：数据管理与挖掘；管理：用数据洞察一切；安全：摆脱大数据风险；平台：信息通信大数据；医疗：数据解决大难题；网络：抓牢数据发源地；零售：打响大数据之战；制造：更快更好地生产；餐饮：精准营销的数据；金融：大数据理财时代；交通：畅通无阻的数据；社会：用数据改变生活。

120个精彩应用案例，图片精美，阐述细致，在学习中找到赚钱商机，从入门到精通大数据！一本在手，轻松玩转大数据，掌握应用与营销，实现从海量到精准，从新手成为大数据应用高手！

本书主要有两个特色：一是容易懂，让抽象的大数据落地到具体行业上；二是接地气，将宏观的大数据与现实相结合，讲解详细，实用性强。

本书细节特色：12大行业领域应用＋15章大数据专题精讲＋110多个经典专家提醒＋120个大数据应用案例＋150多张图片全程图解，帮助读者在最短的时间内掌控大数据的秘密。

适合阅读本书的读者：对数据、数据挖掘、数据分析感兴趣的IT技术人员和决策者，以及实业家、企业高管、营销人员、政府媒体工作人员、创业者、想创业的人和相关专业的学生等。

图书在版编目（CIP）数据

大数据：从海量到精准/李军编著. —北京：清华大学出版社，2014
ISBN 978-7-302-36447-4

I. ①大…　II. ①李…　III. ①数据处理　IV. ①TP274

中国版本图书馆CIP数据核字（2014）第095811号

责任编辑：杜长清
封面设计：刘　超
版式设计：文森时代
责任校对：王　云
责任印制：

出版发行：清华大学出版社
网　　址：http://www.tup.com.cn，http://www.wqbook.com
地　　址：北京清华大学学研大厦A座　　邮　　编：100084
社 总 机：010-62770175　　邮　　购：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质量反馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
印 刷 者：
装 订 者：
经　　销：全国新华书店
开　　本：170mm×230mm　**印　　张：**20　**字　　数：**424千字
版　　次：2014年9月第1版　**印　　次：**2014年9月第1次印刷
印　　数：1～4000
定　　价：48.00元

产品编号：056996-01

前言 Preface

写作驱动

（1）基本概念：大数据是指一般的软件工具难以捕捉、管理和分析的大容量数据，一般以“太字节”（terabyte，TB）为单位。大数据之“大”，并不仅仅在于“容量之大”，更大的意义在于通过对海量数据的交换、整合和分析，发现新的知识，创造新的价值，带来“大知识”、“大科技”、“大价值”和“大发展”，使我们逐渐走向创新社会化的新信息时代。

（2）市场规模：根据 IDC（国际数据公司）的统计，2011 年全球数据总量已经达到 1.8ZB（1ZB 等于 1 万亿 GB，1.8ZB 也就相当于 18 亿个 1TB 移动硬盘的存储量），而这个数值还在以每两年翻一番的速度增长，预计到 2020 年全球将拥有 35ZB 的数据量，增长近 20 倍。据统计，2012 年市场规模达到 4.5 亿元，2014 年还将持续发酵，未来三年有望突破 40 亿元，2016 年有望达到百亿规模。

（3）市场前景：在全球方面，IDC 则预测大数据技术与服务市场将从 2010 年的 32 亿美元攀升至 2015 年的 169 亿美元。在国内，2014 年将是中国供应链大数据快速发展的一年，供应链大数据应用企业必须提前布局占据有力地位。相关调查显示，2013 年中国供应链大数据市场规模将达到 21 亿元，增长率达到 38%，到 2016 年，中国供应链大数据市场规模将达到 59.6 亿元。

（4）应用领域：大数据在企业商业智能、公共服务和市场营销三个领域拥有巨大的应用潜力和商机。今天，大数据似乎成了“万灵药”，从总统竞选到奥斯卡颁奖、从 Web 安全到灾难预测，都能

看到大数据的身影，正如那句俗语："当你手里有了锤子，什么都看上去像钉子"。国内大数据的推广，已经渗透到了公共健康、临床医疗、物联网、社交网站、社会管理、零售业、制造业、汽车保险业、电力行业、博彩业、工业发动机和设备、视频游戏、教育领域、体育领域、电信业等多个行业应用领域。

本书深度结合了国内的大数据发展形势，为读者介绍了简单易行的处理大数据所需的工具、过程和方法，并描绘了一个易于实施的行动计划，以帮助读者发现新的商业机会，实现新的业务流程，做出更明智的决策。

本书特色

最全面的大数据内容介绍：本书集合了大数据的基本概念、基础设施、挖掘方法、风险管理、行业应用等内容，对大数据进行了全面的剖析。

最丰富的大数据案例说明：书中安排了 120 个大数据精彩应用实例，以实例 + 理论的方式，向读者展示了大数据究竟是什么。

最完备的大数据解决方案：书中囊括了各大主流行业的大数据解决方案，通过详尽的分析，让读者看透大数据从海量到精准背后的"魔法"。

本书内容

全书共分为 15 章，具体内容包括入门：大数据的基本概念、价值：大数据商业变革、架构：大数据基础设施、掌握：数据管理与挖掘、管理：用数据洞察一切、安全：摆脱大数据风险、平台：信息通信大数据、医疗：数据解决大难题、网络：抓牢数据发源地、零售：打响大数据之战、制造：更快更好地生产、餐饮：精准营销的数据、金融：大数据理财时代、交通：畅通无阻的数据、社会：用数据改变生活。

适合读者

本书结构清晰、语言简洁，适用于所有对数据、数据挖掘、数据分析感兴趣的 IT 技术人员和决策者阅读，同时也适用于实业家、

企业高管、营销人员、政府媒体工作人员、创业者和想创业的人以及相关专业的学生等学习参考。

作者售后

本书由李军编著，同时参加编写的人员还有：苏高、罗磊、刘嫔、罗林、宋金梅、曾杰、周旭阳、袁淑敏、谭俊杰、徐茜、杨端阳、谭中阳、张国文、李四华、陈国嘉等人。由于时间仓促，加之编者水平有限，书中难免存在疏漏与不妥之处，欢迎广大读者来信咨询和指正，联系邮箱为 itsir@qq.com。

本书声明

本书中所采用的图片、模型等素材，均为所属公司、网站或个人所有，在本书中引用仅为说明之用，绝无侵权之意，特此声明。

编　者



海量数据聚集篇

精准行业聚焦篇



海量数据聚集篇

1

入门：大数据的基本概念

学前提示

互联网的发展带动了云计算、虚拟化、大数据等 IT 新技术的兴起，各行业的互联网化日渐明显，全新 IT 时代正在来临。其中，大数据的兴起和发展成为新 IT 时代行业互联网化最为典型的特征之一。本章将带领读者初步探索大数据的秘密。

要点展示

- 初步认识，大数据究竟是什么
- 预测未来，大数据的发展趋势
- 做好准备，大数据面对的挑战

1.1 初步认识，大数据究竟是什么

随着信息时代的到来，各种数据围绕在我们身边，大数据时代即将到来。但是，很多人并不了解大数据到底是个什么概念。

下面介绍 3 个场景，也许你能从其中找到想要的答案。

【场景 1】：2013 年 4 月 15 日，美国波士顿举行了第 117 届波士顿马拉松大赛，在美东部时间下午 2 时 50 分突然发生两起爆炸，发生地点位于美国马萨诸塞州波士顿科普里广场。爆炸案发生后，美国联邦调查局立即着手调查。波士顿马拉松爆炸案调查部门在 4 月 16 日表示，至少有 1 枚炸弹的制造材料是日常就可购得的压力锅改造而成的，推测可能是国内恐怖分子所为。

2013 年 7 月，在波士顿爆炸案发生 3 个月后，纽约萨克福马县一对夫妻因为妻子用谷歌搜索了“高压锅”，而丈夫在同一时段用谷歌搜索了“背包”。结果，一个由 6 人组成的联合反恐部队，利用“查水表”的名义对这对夫妻进行盘问，“你们有炸弹吗？你们有高压锅吗？为什么只有电饭煲？能拿来做炸弹吗？”

为什么美国政府知道他们有关搜索情况？这一切都归功于“棱镜”和谷歌的数据监视。据悉，类似的上门“查水表”事件，联合反恐部队每周就要进行多达上百次。

由此可见，一个人的搜索信息会成为破案侦查的依据，所以请小心了！

【场景 2】：据某权威机构分析，5 万名手机用户在 3 个月内，无论在家附近活动还是出远门，他们的行踪都相当有规律。一个人大约 93%的行踪在理论上是可预测的。当配偶怀疑对方有了外遇，雇主怀疑雇员把公司的车辆挪为私用，或者是父母想知道他们的孩子是否去了他们所说的那个地方，这些都可以使用如图 1-1 所示的全球卫星定位系统找到所要的地址等信息。

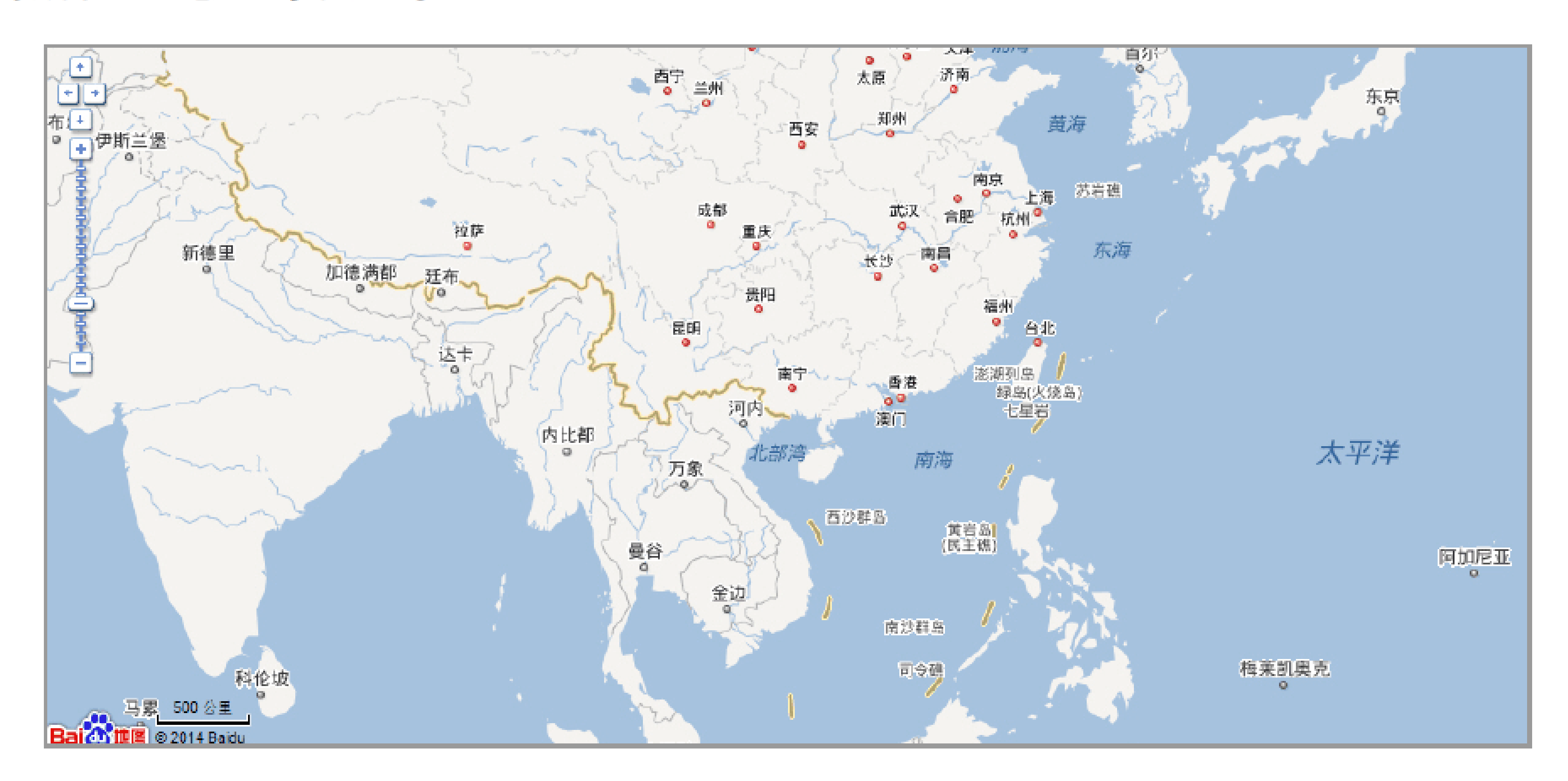


图 1-1　GPS 系统中的地图

利用 GPS 定位系统，再综合多颗卫星的数据，就可以在全球范围内随时找到你或者你的车辆所在的精确位置，如图 1-2 所示。这就是信息、数据时代的威力。





图 1-2 GPS 定位系统可以找到每个人（上图）或者车辆（下图）的精确位置

【场景 3】：2014 年春节，百度推出了“百度迁徙”，其利用大数据技术，对其拥有的 LBS（基于地理位置的服务）大数据进行计算分析，并采用创新的可视化呈现方式，在业界首次实现了全程、动态、即时、直观地展现中国春节前后人口大迁徙的轨迹与特征，如图 1-3 所示。查询网址：http://qianxi.baidu.com/。

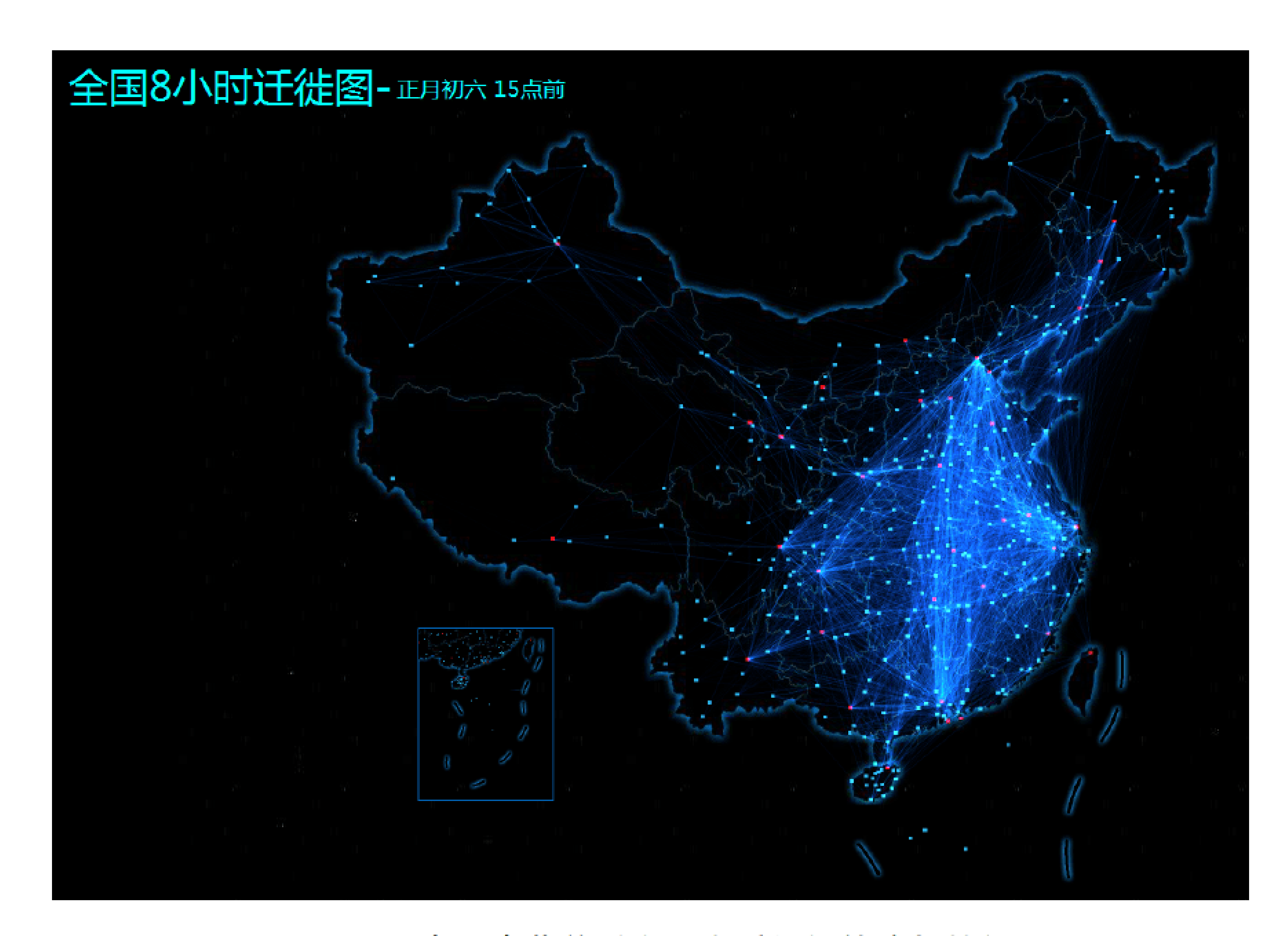


图 1-3　中国春节前后人口大迁徙的轨迹与特征

用户还可以查询某一个城市的“迁入城市”、“迁出城市”的最新数据迁徙图，如查询“北京”的迁徙情况，如图 1-4 所示。



图 1-4　春节期间北京的迁徙情况

1.1.1　大数据基本定义

前面洋洋洒洒地说了很多，相信很多读者看到过相关的报告，但是截至目前，我们始终没有给出大数据的定义，也就是说我们并没有清楚地表述过：大数据到底是什么。

在 IDC（Internet Data Center，互联网数据中心）的报告中，他们对大数据进行了一个简单的描述：Big data is a big dynamic that seemed to appear from nowhere. But in reality, big data isn't new. Instead, it is something that is moving into the mainstream and getting big attention, and for good reason. Big data is not a "thing" but instead a dynamic/activity that crosses many IT borders。

中文翻译为：大数据是一个看起来似乎来路不明的大的动态过程。但实际上，大数据并不是一个新生事物，虽然它确确实实正在走向主流和引起广泛的注意。大数据并不是一个实体，而是一个横跨很多 IT 边界的动态活动。如图 1-5 所示为 IDC 所描述的大数据世界。



图 1-5　IDC 所描述的大数据世界（资料来源：IDC）

如果 IDC 的解释也能算是大数据的一种描述性定义的话，相信大部分人应该是很难理解大数据的。

因此，想要明白"大数据"的概念，还要从"大数据"的名词本身入手。首先要从"大"入手，那么"大数据"的"大"到底指的是哪些方面呢？笔者认为，大数据同过去的海量数据有所区别，其基本特征可以用 4 个 V 来总结（Volume、Variety、Value 和 Velocity），即体量大、多样性、价值密度低、速度快。

- 数据体量大：大数据一般指在 10TB 规模以上的数据量。但在实际应用中，很多企业用户把多个数据集放在一起，已经形成了 PB 级的数据量。
- 数据多样性：数据来自多种数据源，数据种类和格式日渐丰富，已经冲破了以前所限定的结构化数据范畴，囊括了半结构化和非结构化数据。
- 价值密度低：大数据所创造的价值密度明显更低。根据福利经济学的观点，生产率与单位商品的价值无关，生产率只与生产的数量有关，即生产率高的企业在相同的时间内生产更多的价值——因而可以把更高的生产率理解为通过生产和管理技术的革新而形成的更高的劳动复杂度，劳动复杂度的提高使单位劳动时间具有了更大的价值密度。

➢ 速度快：有数据显示，在全球范围内，数据量以每年 50%的速度增长，数据增长的速度已经远远超过 IT 设计发展的速度。数据本身已经成为企业发展的资产。快速捕捉数据信息，实现数字化生产和管理，已经成为未来企业赢得市场，应对行业互联网化的必经之路。

另外，从“数据”这个词来分析，大数据是海量的，是巨大的，它关乎数据量。笔者认为可以从 3 个方面定义大数据：（1）数据量；（2）广度、分类；（3）速度。简而言之，大数据就是一个体量特别大，数据类别特别丰富的数据集。也就是说“大数据”本身并不是一种新的技术，也不是一种新的产品，而是我们这个时代出现的一种现象。而这个“大”大到了一种什么样的程度呢？可以说它即将突破现有常规软件所能提供的能力极限。

综上所述，全球最大的战略咨询公司麦肯锡给出了一个十分明确的定义：大数据是指无法在一定时间内用传统数据库软件工具对其内容进行抓取、管理和处理的数据集合。

随着互联网革命性地改变了商业的运作模式、政府的管理方法以及人们的生活方式，信息的积累足以引发新的变革。世界充斥着比以往更多的信息，信息总量的变化导致了信息形态的变化。“大数据”这一概念应运而生。“大数据”不同于互联网，它正在以巨大的力量改变着世界，它是具有更强的决策力、洞察力、流程优化能力、高增长率和多样化的信息资产。

如今，数据库、大数据已经成为变革的中心，事实上可以成为一场革命。在 IT 领域、制造业、零售业、政府管理、科技领域，大数据都在改变着这个世界的运行方式。因此，我们称之为大数据的新世界。

专家提醒

数据基本单位换算：

1B（byte，字节）= 8b（bit 位）

1KB（Kilobyte，千字节）= 1024B

1MB（Megabyte，百万字节兆字节，简称“兆”）= 1024KB

1GB（Gigabyte，十亿字节吉字节，又称“千兆”）= 1024MB

1TB（Trillionbyte，万亿字节太字节）= 1024GB

1PB（Petabyte，千万亿字节拍字节）= 1024TB

1EB（Exabyte，百亿亿字节艾字节）= 1024PB

1ZB（Zettabyte，十万亿亿字节泽字节）= 1024EB

1.1.2 大数据结构特征

如今，全球存储的数据量正在急剧增长，数据量大是大数据的一致特征。在 2000 年，全球存储了 800000PB 的数据。预计到 2020 年，这一数字会达到 35ZB。单单 Twitter 每天就会生成超过 7TB 的数据，Facebook 为 10TB，一些企业在一年中每一天的每一

小时就会产生数 TB 的数据。

就传统 IT 企业来看，其结构化和非结构化的数据增长也是惊人的。2005 年企业存储的结构化数据为 4EB，到 2015 年将增至 29EB，年复合增长率逾 20%。非结构化数据发展更猛。2005 年为 22EB，2015 年将增至 1600EB，年复合增长率约 60%，远远快于摩尔定律。

那么，一分钟到底会有多少数据产生呢?

- 电子邮件用户发送 204166677 条信息。
- Google 收到超过 2000000 个搜索查询。
- Facebook 用户分享 684478 条内容。
- 消费者在网购上花费 272070 美元。
- Twitter 用户发送超过 100000 条微博。
- 苹果公司收到大约 47000 个应用下载。
- Facebook 上的品牌和企业收到 34722 个 “赞”。
- Tumblr 博客用户发布 27778 个新帖子。
- Instagram 用户分享 36000 张新照片。
- Flickr 用户添加 3125 张新照片。
- Foursquare 用户执行 2083 次签到。
- 571 个新网站诞生。
- WordPress 用户发布 347 篇新博文。

由于数据自身的复杂性，作为一个必然的结果，处理大数据的首选方法就是在并行计算的环境中进行大规模并行处理（Massively Parallel Processing，MPP），这使得并行摄取、并行数据装载和分析成为可能。实际上，大多数的大数据都是非结构化或者半结构化的，这需要不同的技术和工具来处理和分析。

大数据的结构就体现了它最突出的特征，如表 1-1 所示，显示了几种不同数据结构类型数据的增长趋势。据悉，未来数据增长的 80%～90%将来自于非结构化的数据类型（包括半非结构化、准非结构化和非结构化数据）。

表 1-1　数据增长日益趋向非结构化

结构化进程	数 据 内 容	举　　例
结构化	包括预定义的数据类型、格式和结构的数据	事务性数据和联机分析处理
半结构化	具有可识别的模式并可以解析的文本数据文件	自描述和具有定义模式的 XML 数据文件
准结构化	具有不规则数据格式的文本数据，通过使用工具可以使之格式化	包含不一致的数据值和格式的网站点击数据
非结构化	没有固定结构的数据，通常将其保存成不同类型的文档	TXT 文本文档、PDF 文档、图像和视频

1.1.3 大数据与云计算

在过去 3 年当中，笔者经历了大数据的发展从无到有，3 年前可能还没有人说这个词，现在已经如火如荼。现在，每天有大量数据和信息生成，这为大数据分析提供了机会。相较于传统数据，大数据更能反映这个世界的真实情况，例如，人们会上传和公布大量的图片来记录个人的生活和社会的变化。如今，一天之内人们上传的照片数量就相当于柯达发明胶卷之后拍摄的图像总和。

过去，计算机主要是用于解决大企业交易型的数据，并不会记录其他无关的信息，只有在云计算产业规模化发展之后，分布式计算才给大数据提供了记录的载体。可以说，云计算使大数据变成可能，打个比方，云计算充当了工业革命时期“发动机”的角色，而大数据则是“电”。

然而，现在除了数据本身发生了改变，云计算也使数据变得更加分散，在这样的趋势下，传统数据库对于海量数据存储的需求、处理速度的需求、数据多样化的需求难以满足，从而使各种各样的解决方案大行其道。

总之，云计算为大数据带来了硬件存储的条件——更便宜的分布式运算存储，而互联网时代的今天也在不断呼唤数据应用和服务。在技术和需求的双重推动下，会有越来越多的政府机构、公司企业和个人意识到数据是巨大的经济资产，像货币或黄金一样，它将带来全新的创业方向、商业模式和投资机会。

大数据和云计算的区别与联系如表 1-2 所示。

表 1-2　大数据和云计算的区别与联系

具体表现	区　别	联　系
概念	云计算改变了 IT，而大数据则改变了业务	大数据必须有云作为基础架构，才能得以顺畅运营
目标受众	云计算是卖给 CIO 的技术和产品，是一个进阶的 IT 解决方案；大数据是卖给 CEO、业务层的产品，大数据的决策者是业务层	由于它们能直接感受到来自市场竞争的压力，因而必须在业务上以更有竞争力的方式战胜对手

专家提醒

云计算和大数据注定将带来一次革命，无论是对社会、公司和个人来说，都是一次世界观的改变。届时，互联网不再是一个展示公司的工具或平台，而是属于未来的生产方式，是关乎竞争和生存的关键。

1.1.4 大数据规模预测

当你走进一家陌生的小餐厅时，耳边响起只有你才熟悉的音乐旋律。这样的场景实

现技术上并不难，餐厅只要读出你的手机音乐下载记录，通过数据分析，就可以定制播放你喜欢的音乐，这就是大数据时代的潜力。

前面笔者已经说了，大数据由 4 个 V 组成，这 4 个 V 的组合推动了第 5 个因素——价值（Value）的出现。随着云计算概念日渐深入人心，大数据也越来越受到关注。国际知名数据公司 IDC 在长期对云计算市场进行跟踪研究的同时，也对大数据市场保持着密切关注。如图 1-6 所示，IDC 发现，目前大数据对市场的影响正日益提升，已经开始影响数据中心设计、移动应用投资、数据管理等相关领域。

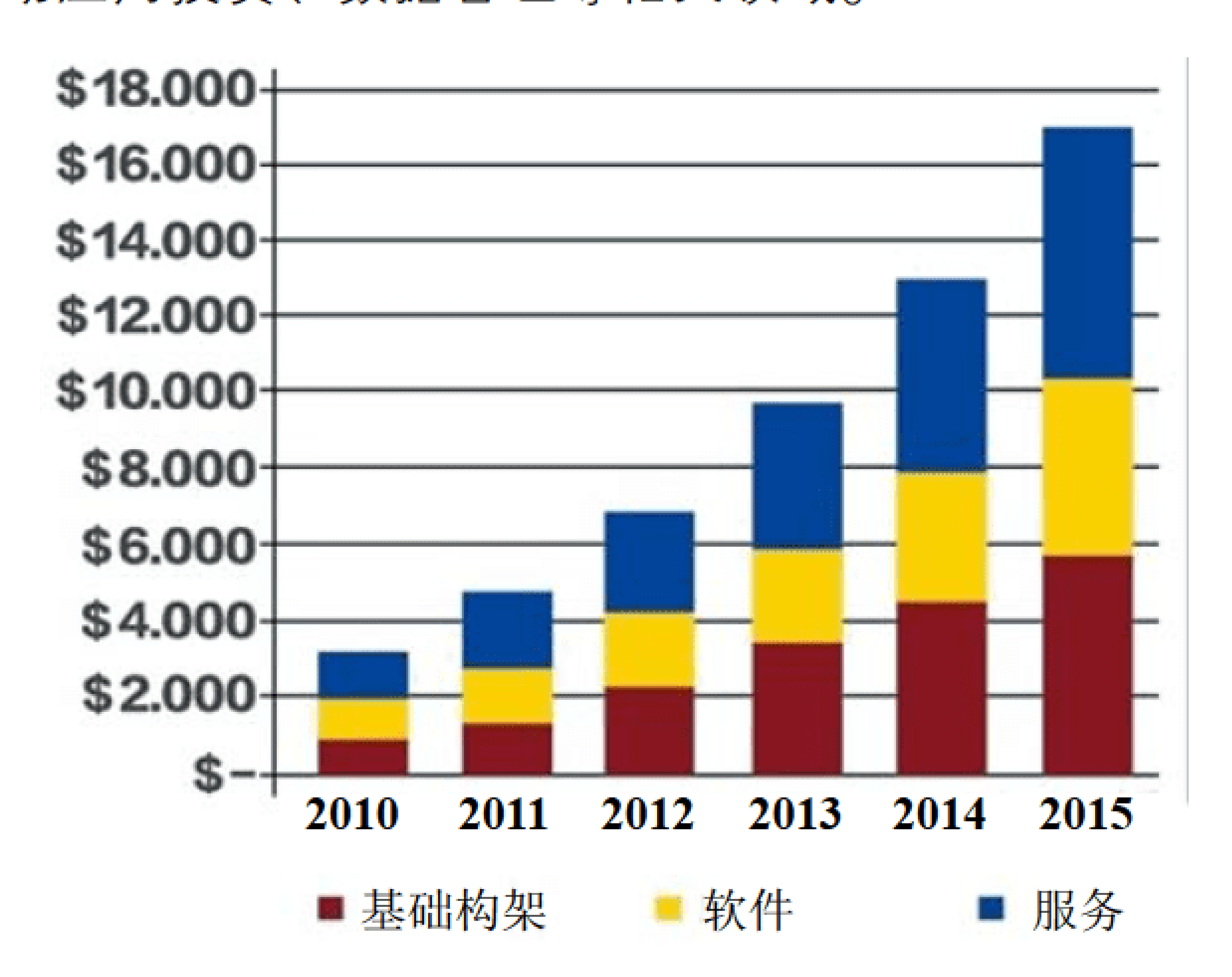


图 1-6　IDC 全球大数据市场规模与预测

1.1.5　大数据的发展史

如今，越来越多的企业参与到大数据的竞争中来，那么“大数据”这个词汇是如何诞生以及演变的呢?

大数据是一个修辞学意义上的词汇，在数据方面，“大”（big）是一个快速发展的术语。早在 1890 年，美国统计学家赫尔曼・霍尔瑞斯为了统计这一年的人口普查数据，发明了一台电动器来读取卡片上的数据，该设备让美国用一年时间就完成了原本耗时 8 年的人口普查活动，由此在全球范围内引发了数据处理的新纪元。

1961 年，刚成立 9 年的美国国家安全局（NSA）是拥有超过 12000 个密码学家的情报机构，在间谍饱和的冷战年代，面对超量信息，他们开始采用计算机自动收集处理信号情报，并努力将仓库内积压的模拟磁带信息进行数字化处理。仅 1961 年 7 月份，该机构就收到了 17000 卷磁带。

起初，许多科学家和工程师都嘲笑“大数据”只不过是一个营销术语。2008 年末，“大数据”得到部分美国知名计算机科学研究人员的认可，业界组织“计算社区联盟”（Computing Community Consortium）发表了一份有影响力的白皮书《大数据计算》，

中肯地阐述了大数据带来的机遇和挑战。

2009 年 5 月，美国总统巴拉克・奥巴马政府推出 data.gov 网站，作为政府开放数据计划的部分举措。该网站拥有超过 4.45 万的数据量集，这样一些网站和智能手机应用程序能跟踪如航班、产品召回、特定区域内失业率等信息，这一行动激发了肯尼亚、英国等政府相继推出类似举措。

2011 年 2 月，扫描 2 亿页的页面信息，或 4 兆兆字节磁盘存储，只需几秒即可完成。同时，IBM 的沃森计算机系统在智力竞赛节目《危险边缘》中打败了两名人类挑战者，后来《纽约时报》称这一刻为“大数据计算胜利”的时刻。

2011 年，英国《自然》杂志曾出版专刊指出，倘若能够更有效地组织和使用大数据，人类将得到更多的机会发挥科学技术，这对社会发展有巨大的推动作用。

2012 年 3 月，美国政府报告要求每个联邦机构都要有一个“大数据”的策略，作为回应，奥巴马政府宣布了一项耗资两亿美元的大数据研究与发展项目。

2012 年 7 月，美国国务卿希拉里・克林顿宣布了一个名为“数据 2X”的公私合营企业，用来收集统计世界各地的妇女和女童在经济、政治和社会地位方面的信息。

回顾过去的 50 多年，我们可以看到 IT 产业已经经历了几轮新兴和重叠的技术浪潮，如图 1-7 所示。这里面的每一波浪潮都是由新兴的 IT 供应商主导的，他们改变了已有的秩序，重新定义了已有的计算机规范，并为进入新时代铺平了道路。

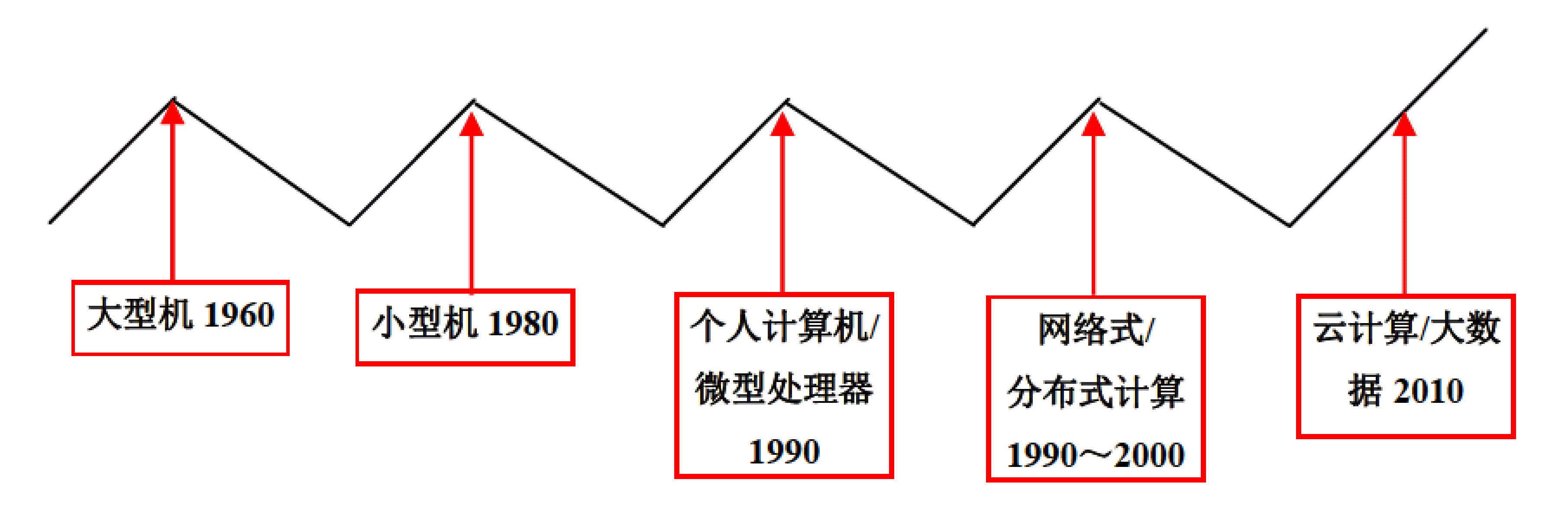


图 1-7　IT 产业的发展浪潮

人们手中的手机和移动设备是数据量爆炸的一个重要原因，目前，全球拥有 50 亿台手机用户，其中 20 亿台为智能电话，这相当于 20 世纪 80 年代 20 亿台 IBM 的大型机掌握在消费者手里。

“大数据”是“数据化”趋势下的必然产物。数据化最核心的理念是：“一切都被记录，一切都被数字化”。它带来了两个重大的变化：一是数据量的爆炸性剧增，最近两年所产生的数据量等同于 2010 年以前整个人类文明产生的数据量总和；二是数据来源的极大丰富，形成了多源异构的数据形态，其中非结构化数据所占比重逐年增大。

1.1.6　大数据技术架构

即便是在“摩尔定律”，即每 18 个月芯片性能将提高 1 倍的支撑下，硬件性能进化

的速度也早已赶不上数据增长的速度了，并且差距越来越巨大。例如，一分钟之内，新浪微博有数万条微博发送，苹果应用商店下载次数以万计，淘宝卖出了几万件商品，百度产生了百万次搜索查询……所有这些行为都由海量的数据来呈现。

那么，大数据是通过什么样的技术架构来接受、容纳并处理这些海量数据的呢?

要容纳数据本身，IT 基础架构必须能够以经济的方式存储比以往更大量、类型更多的数据。此外，还必须能适应数据速度，即数据变化的速度。数量如此大的数据难以在当今的网络连接条件下快速来回移动。大数据基础架构必须具有分布式计算能力，以便能在接近用户的位置进行数据分析，减少跨越网络所引起的延迟。

因此，云计算模式为大数据的成功提供了很好的条件，以实现大数据分析所需的效率、可扩展性、数据便携性和经济性。另外，还可以用来跨越毫不相干的数据源比较不同类型的数据和进行模式匹配。这使得大数据分析能以新视角挖掘企业传统数据，并带来传统上未曾有过的数据洞察力。

例如，LinkedIn 是世界上最大的专业人士社交网络，在全球范围内有 2.25 亿用户，并且以每秒 2 个新用户的速度增长。LinkedIn 还是一个解决方案供应商，据悉，目前有 88%的财富 100 强企业在使用 LinkedIn 的付费解决方案，LinkedIn 还有超出 290 万的公司主页及相关信息。

LinkedIn 之所以取得如此大的成功，是因为他们有专业的身份可以拓展人脉发现机遇，专业的内容全方位掌握业界资讯，专业的平台随时随地了解人脉动向。

从 LinkedIn 的业务模型不难看出，其本身就拥有海量的数据，通过这些数据创造出有价值的产品和服务，来增加用户数量和用户黏性，这样数据还会不断增长从而形成一个“闭环”。LinkedIn 有人才、市场、高级订阅服务三大商业解决方案，而且三大商业解决方案的盈收每年也呈翻倍增长趋势，而其中占盈收比例最大的是人才解决方案。

另外，LinkedIn 的数据按用户可分为用户特征数据、用户行为数据、用户网络数据；按数据存取速度可分为在线数据、近线数据、离线数据。LinkedIn 的三级数据架构根据不同性质的工作设计，其中近线数据存储在 Voldemort 分布式数据库中，在线数据存储在 Oracle 和 Espresso 中，服务器日志存储在 Web Logs 中。使用 Kafka 发布数据，通过 Databus 捕获在线数据，而所有的离线数据由 Hadoop 和 Teradata 数据库构成。

基于上述考虑，大数据可以采用四层堆栈式技术架构，如表 1-3 所示。

表 1-3　采用四层堆栈式技术架构的大数据

层　次	说　明	特　点	作　用
基础层	第一层作为整个大数据技术架构基础的最底层，也是基础层	虚拟化、网络化、分布式；横向可扩展体系结构	要实现大数据规模的应用，企业需要一个高度自动化的、可横向扩展的存储和计算平台。这个基础设施需要从以前的存储孤岛发展为具有共享能力的高容量存储池。容量、性能和吞吐量必须可以线性扩展

续表

层　　次	说　　明	特　　点	作　　用
管理层	本层既包括数据的存储和管理，也涉及数据的计算	处理结构化数据和非结构化数据；并行处理，线性可扩展	由于并行化和分布式是大数据管理平台所必须考虑的要素，因此要支持在多源数据上做深层次的分析，大数据技术架构中需要一个管理平台，使结构化和非结构化数据可一体化管理，具备实时传送和查询、计算功能
分析层	大数据应用需要大数据分析	提供自助服务；使用灵活，实时协作	分析层提供基于统计学的数据挖掘和机器学习算法，用于分析和解释数据集，帮助企业获得对数据价值深入的领悟。可扩展性强、使用灵活的大数据分析平台更可成为数据科学家的利器，从而达到事半功倍的效果
应用层	大数据的价值体现在帮助企业进行决策，以及为终端用户提供服务应用	提供实时决策，内置预测能力；利用数据驱动经济，使数据实现货币化	不同的新型商业需求驱动了大数据的应用。反之，大数据应用为企业提供的竞争优势使得企业更加重视大数据的价值。新型大数据应用对大数据技术不断提出新的要求，大数据技术也因此在不断地发展变化中日趋成熟

专家提醒

云模型鼓励访问数据并提供弹性资源池来应对大规模问题，其解决了如何存储大量数据，以及如何积聚所需的计算资源来操作数据的问题。在云中，数据可跨多个节点调配和分布，这使得数据更接近需要它的用户，从而缩短响应时间和提高生产率。

1.1.7　大数据重要的理由

人们为什么如此关心大数据呢？其实大数据可以使我们提出新问题，来了解我们的业务。例如社交网络分析，一个企业，即使你是一个个体，你也有一个品牌，如何分析你的品牌影响力、品牌声誉，这些问题之前不容易回答，如今在大数据的时代可以很容易得到答案，并且几乎是以实时的速度来解答。

例如，有一家物流公司，有卡车等运输工具，希望优化车队的运输路线，提高运输效率，并且基于实时的交送信息、天气信息及其他类型的信息。现在通过传感器和大数据就可以做到。事实上，关于过去和现在，甚至是未来的事务，大数据分析都能够用得上。

专家提醒

虽然大数据是一个重大问题，但笔者认为，真正的问题是如何让大数据更有意义，如何在大数据里面寻找模式帮助组织机构做出更好的商业决策。

当前，随着互联网科技的日益成熟，各种类型数据的增长将会超越历史上任何一个时期。因此，用户想要从这庞大的数据库中提取对自己有用的信息，就离不开大数据分析技术和工具。如表 1-4 所示，向大家展示了大数据分析将越来越重要的 10 个理由。

表 1-4　大数据分析为何重要的理由

理　由	说　明
Hadoop 用户迅速增长	越来越多的企业开始使用 Hadoop 平台处理大量数据。例如，2009 年 Hadoop 服务提供商总共只有 9 家，而在 2012 年就已经超过了 120 家
Hadoop 整合功能加深	仅靠 Hadoop 服务是无法解决企业的大数据问题的，很多传统的数据库管理系统开始整合 Hadoop 服务，以便更好地为企业服务。例如，惠普、戴尔、甲骨文、IBM 等知名公司都分别有针对自家需求的 Hadoop 服务
更多 Hadoop 服务走上云端	云端上的 Hadoop 服务让大数据分析和处理更加方便快捷
原始数据的价值	在相关大数据分析处理技术出现之前，IT 公司经理们通常要对公司数据进行筛选以便用户查询和分析，现在，各种大数据分析工具既方便用户查询分析数据，又能避免泄露公司机密，同时，所有原始数据都将完好保存
大数据开发技术的“短板”得以解决	阻碍大数据分析技术或是使用 Hadoop 的原因之一就是缺乏相应的技术、环境、数据安全以及可行性。幸好，许多开源和专利软件社区都已经着手解决这些问题了，使大数据的“短板”逐渐消失
ROI 案例分析将成为主流	许多传统企业（包括银行、电信公司和零售商）都开始使用 Hadoop 服务，但很少有人愿意分享所有细节，所以很难找出一个真正的 ROI（投资回报率）案例进行分析，这促使大数据分析势在必行
其他大数据分析平台的兴起	一说到大数据，很多人第一时间想到的就是 Hadoop，其实还有许多其他不错的大数据分析平台，如 Platfora、Datahero 等
磁盘终将被历史淘汰	目前，应该有一半以上的企业还在利用磁盘进行数据存档、备份和恢复。但随着大数据分析技术日渐成熟，磁盘终将被淘汰
机器学习和人工智能的崛起	机器学习和人工智能正在崛起，但在银行、金融服务、电信以及制造等传统行业它们仍是十分稚嫩的新兴技术
Hadoop 将继续发展	Hadoop 仍处在初级阶段，未来还将具备更多功能，例如，自由文本搜索功能以及基于 GUI（图形用户界面）的可视化工具

专家提醒

对大企业而言，大数据的兴起，首先，是因为计算能力可以更低的成本获得，且各类系统如今已能够支持多任务处理；其次，内存的成本也在直线下降，企业可以在内存中处理比以往更多的数据；最后，把计算机聚合成服务器集群越来越简单。

1.1.8 大数据的解决方案

当前，越来越多的企业将大数据的分析结果作为其判断未来发展的依据。同时，传统的商业预测逻辑正日益被新的大数据预测所取代。既然大数据如此重要，那么大数据解决方案是否可以完全替代传统的数据库解决方案呢?

在这里，笔者先不说出答案，而是先带大家看一个典型的案例：

例如，一个优秀的棒球运动员知道自己的哪一只手更擅长抛球，哪一只手更擅长接球。就像这样一种情形，每只手可以尝试执行它天生不适合的任务，但会非常笨拙，因此，通常不会看到棒球运动员使用一只手接球，停下来，丢掉他们的手套，然后使用同一只手抛球。棒球运动员的左手和右手协同起来会实现最佳的结果。

上面的例子就是传统数据库和大数据技术的一个简单类比：没有这两个重要实体的协同工作，任何组织或结构的信息平台都很难得到进一步发展，因为就像棒球运动员协调双手来抛接棒球一样，一个团结一致的分析生态系统才能实现最佳的结果。

此时，我们经过初步分析就可以了解到，有些类型的问题不是本来就属于传统数据库的，至少在最初不是，而且也不确定是否希望将一些数据放在仓库中，因为我们不知道它是否拥有较高的价值、是否是非结构化的，或者是否太庞大了。更多的情况是，在投入精力和金钱将数据放在仓库之后，才能发现每个字节的数据价值；但我们希望在投资之前，就能明确该数据值得保存，并拥有较高的价值。

典型的大数据解决方案应该是具有多种能力的平台化解决方案，这些能力包括结构化数据的存储、计算、分析和挖掘，多结构化数据的存储、加工和处理，以及大数据的商务智能分析。笔者认为，这种解决方案在技术上应具有以下 4 个特性：软硬集成化的大数据处理能力、全结构化数据处理的能力、大规模内存计算的能力、超高网络速度访问的能力。

因此，你一定要认识到传统数据库技术是整体解决方案中一个重要且相关的部分。事实上，它们在与你的大数据平台结合使用时会变得更加重要。

专家提醒

当前，越来越多的企业将大数据的分析结果作为其判断未来发展的依据。同时，传统的商业预测逻辑正日益被新的大数据预测所取代。但是，笔者觉得大家对于大数据的期望值要谨慎一些，因为海量数据只有在得到有效治理的前提下，才能进一步发挥其价值。

1.2 预测未来，大数据的发展趋势

据悉，在 1993 年的美国《纽约人》杂志上刊登了一幅标题为“互联网上，没有人

知道你是一条狗”的漫画，而作者彼得·施泰纳也因此赚取了超过 5 万美元。此后的 20 年间，互联网发生了巨大的变化，移动互联、社交网络及电子商务大大拓展了互联网的疆界和应用领域。

如今，我们在享受便利的同时，也无偿贡献了自己的“行踪”，现在互联网不但知道对面是一只狗，甚至还知道这只狗喜欢什么食物，几点出去遛弯，几点回窝睡觉。每个人在互联网进入到大数据时代，都将是透明性存在的，可以说是“处处行迹处处留痕”。

收集并分析海量的各种类型数据，并快速获取影响未来的信息的能力，这就是大数据技术的魅力。事实上大数据的来源非常广泛，天上的卫星、地上的汽车、埋在土壤里面的各类传感器，无时无刻不在生成大量的数据。这些数据如果加以综合利用，产生的社会价值和经济价值将是难以估量的。大数据技术让人们看到未来解决预测问题的一丝曙光。

1.2.1 大数据撬动全世界

大数据不仅体现为数据量的惊人增长，更前所未有地引入了正在不断扩展中的数据类型。从量的增长来看，根据 IDC（国际数据公司）的跟踪分析，全球产生的数据总量 2011 年已经达到 1.8ZB（1ZB 等于 1 万亿 GB，1.8ZB 也就相当于 18 亿个 1TB 移动硬盘的存储量）；2012 年达到约 2.8ZB，但当年全球产生的数据中仅有约 0.5%得到有效分析。据悉，到 2020 年，全球数据总量中有 22%将来自中国。

电商投放广告、物流调度运力、证监会抓老鼠仓、金融机构卖基金、民航节约成本、农民破解猪周期、制片人拍电影……看似毫不相关的事情，背后都有大数据在发力。随着互联网、移动互联网对各个领域的渗透越来越深，从政府到企业，从群体到个人，数据的积累与日俱增。4G 牌照的发放，又让移动数据通道由“乡村公路”升级为“高速公路”。

与此同时，社会上的各行各业，从电信、IT 业，到金融、证券、保险、航空、酒店服务业等，地球上的各种存在事物，从每个人到每棵树、每朵花乃至每粒沙子，无一例外地都在成为大数据的生成者。笔者可以预见，大数据席卷各行各业和人们生活的速度只会越来越快。

例如，世界上第一部“先拍照后对焦”光场相机 Lytro，就运用了大数据处理分析理念。与传统相机只记录一束光不同，Lytro 可以记录整个光场里所有的光，也就是用总体数据取代了随机样本。用户没必要一开始就对焦，想要什么样的照片可以在拍摄之后再决定。

因此，究竟该如何“开采”大数据这座丰富的矿藏，成为了一个令人着迷的问题，因为与正确答案相随的将是谁都渴望的巨大商业成功。当前，伴随着变革的发生，传统的互联网企业已经站在了大数据时代的最前沿。作为后 PC 时代的 4 大巨头，Facebook、

谷歌、苹果、亚马逊正在成为大数据的拥有者和使用者，其主要特点如表 1-5 所示。

表 1-5　4 大互联网企业的大数据策略

互联网企业	大数据策略
Facebook	依靠其强大的社交网络，已然成为业界第一个生成大数据的“巨鳄”
苹果	依靠操作系统和颠覆性的终端，正在努力打造大数据的生成之地
谷歌	主要依靠操作系统、搜索引擎和“Google+”平台整合终端产品，以储备可以利用的大数据
亚马逊	作为云计算的最早倡导者之一，则通过网络平台、云计算平台和阅读终端，期望建立起一个电子商务垂直领域的大数据汇集地

大数据，正在撬动全世界的神经，无论是国家、企业，还是每一个独立存在的个人，都将成为大数据时代的贡献者和受益者。

专家提醒

目前，数据量的大幅增加对人们注重精确性的习惯提出了挑战。大数据需要技术和思维上的变革才能利用，才能做到从海量到精准。这一轮的变革，事关绝大多数企业的命运。可以看到，用大数据这个视角，可以考察企业的兴衰。第一，如果对大数据不关心，不了解，必将走向衰败；第二，拥有大量的数据并善加运用的公司，必将赢得未来。时代变了，判断企业价值的标准、判断软件价值的标准也变了。

1.2.2　大数据是大势所趋

大数据有多火？有媒体将 2013 年称为“大数据元年”。目前，几乎所有世界级的互联网企业，都将业务触角延伸至大数据产业；无论是社交平台逐鹿、电商价格大战还是门户网站竞争，都有它的影子。2012 年，美国政府投资两亿美元启动“大数据研究和发展计划”，更将大数据上升到国家战略层面。大数据，正在由技术热词转变为一股社会浪潮，影响社会生活的方方面面。

星巴克有意推出的“大数据咖啡杯”就是个小小的例子。美国媒体报道，这家咖啡连锁巨头打算试验在一些咖啡杯中装上传感器，收集常客喝咖啡速度等数据，从而为喝咖啡较慢顾客提供保温效果好的杯子，以提高其满意度和忠诚度。

又例如，在 2008 年初，阿里巴巴平台上整个买家询盘数急剧下滑，欧美对中国采购量也在下滑。通常而言，买家在采购商品前，会比较多家供应商的产品，反映到阿里巴巴网站统计数据中，就是查询点击的数量和购买点击的数量会保持一个相对的数值。

阿里巴巴平台通过统计历史上所有买家、卖家的询价和成交的数据，可以形成询盘指数和成交指数。这两个指数是密切相关的：询盘指数是前兆性的，前期询盘指数活跃，就会保证后期一定的成交量。因此，当马云观察到询盘指数异乎寻常地下降，自然就可

以推测未来成交量的萎缩。这种统计和分析，如果缺少大数据技术的支持，是难以完成的。这次事件，马云得以提前呼吁，帮助成千上万的中小制造商准备“过冬粮”，从而赢得了很高的声誉。

因此，大数据是一种新的价值观和方法论，人们面对的不再是随机样本而是全体数据，不是精确性而是混杂性，不是因果关系而是相关关系。

1.2.3 大数据将成为资产

众所周知，用户的消费习惯、兴趣爱好、关系网络以及整个互联网的趋势、潮流都将成为互联网从业者关注的热点，而这一切的获取和分析都离不开大数据，因为在社会化媒体基础上的大数据挖掘和分析都会衍生很多应用。例如，帮企业做内部数据挖掘，帮企业找到更精准用户，降低营销成本，提高企业销售率，增加利润等。

大数据、社会化媒体营销真正实现了营销模式的“量体裁衣”，这是营销领域跨时代的进步。未来企业的竞争，将是拥有数据规模和活性的竞争，将是对数据解释和运用的竞争。

随着技术的发展，大数据社会化营销将是未来营销的主战场，即将到来的大数据时代可以在任何行业，任何服务上出现，由此可能产生的服务和商业模式将是无穷尽的。笔者认为，围绕大数据至少可以演绎出 6 种新的商业模式，如表 1-6 所示。

表 1-6 6 种新的商业模式

商业模式	主要特点
出租或出售数据	即通过出售广泛收集、精心过滤时效性强的数据来获得收益，这也是“数据就是资产”的最经典诠释
出租或出售信息	需要注意的是，这里的信息指的是经过加工处理，承载一定行业特征的数据集合。一般来讲聚焦某个行业，广泛收集相关数据，深度整合萃取信息，以庞大的数据中心加上专用传播渠道，也可取得成功
数字媒体精准营销	这个模式最性感，因为全球广告市场空间是 5000 亿美元，具备培育千亿级公司的土壤和成长空间。这类公司的核心资源是获得实时、海量、有效的数据，立身之本是大数据分析技术，盈利来源是精准营销
数据分析业务	该模式令人着迷之处在于，如果没有大量的数据，缺乏有效的数据分析技术，这些公司的业务其实难以开展。例如，以阿里金融为代表的小额信贷公司，通过在线分析小微型企业的交易数据、财务数据，甚至可以计算出应提供多少贷款，多长时间可以收回等关键问题，把坏账风险降到最低
运营数据空间	传统的 IDC 和互联网巨头们都在提供此类服务，而且其他 IT 企业也纷纷嗅到了大数据的商机，开始抢占个人、企业的数据资源。海外的 Dropbox，国内微盘都是此类公司的代表。这类公司的想象空间是它可以成长为数据聚合平台，盈利模式将趋于多元化

续表

商 业 模 式	主 要 特 点
大数据处理业务	从数据量上来看，非结构化数据是结构化数据的 5 倍以上，任何一个种类的非结构化数据处理，都可以重现现有结构化数据的辉煌。语音数据处理领域、视频数据处理领域、语义识别领域、图像数据处理领域都可能出现大型的、高速成长的公司

如今，“大数据”这一话题在国内受到投资者追捧，也不断有高技术人才选择这个方向创业；但实际上国外对于“大数据”，已经走过了概念炒作阶段，进入到实际的应用，产生了实际的效益。例如，美国奥巴马政府已经开始大规模地投资大数据领域，这是大数据从商业行为上升到国家战略的分水岭，表明大数据正式提升到战略层面，大数据在经济社会各个层面、各个领域都开始受到重视。笔者相信，“大数据”将领跑新一轮互联网投资高潮，让资产逐步变成资本。

1.2.4 大数据时代的转变

互联网的重心逐步向着移动互联转移，各种新型智能移动设备的迅速普及带来了海量数据的爆发。于是大家都在谈论大数据，大家都想用好大数据。但你真的了解大数据吗？当前的行业状况又是怎样？

事实上，大数据只是一种提法，其形态本身是数据云。因此，以实时感知、分析、对话、服务能力为基础，让数据流成为商业、营销活动的核心才是关键。怎样才能让这些大数据更好地为产品或营销服务，搞清楚大数据时代的业界生态必不可少。

我们可以结合互联网数据中心（Data Center of China Internet，DCCI）发布的数据报告一起来看看。

1. 互联网生态结构：传统互联网→移动互联网

据市场研究机构 IDC 预测，2013 年全球智能手机出货量将超过 10 亿部，这个数字意味着它比 2012 年增长了近 40%。

同时关于三大移动智能操作系统，我们还得到这样一组数据，如表 1-7 所示。

表 1-7 三大移动智能操作系统的 APP 相关数据

操 作 系 统	APP 商店	上 线 时 间	主 要 数 据
iOS	Apple App Store	2008 年 7 月 11 日	App 数量：65 万余款
			下载数量：300 亿次
			设备激活总量：3.65 亿
Android	Google Play Market	2008 年 10 月 22 日	App 数量：60 万余款
			下载数量：200 亿次
			设备激活总量：4 亿

续表

操作系统	APP 商店	上线时间	主要数据
Windows	Widows Phone Marketplace	2010 年 10 月 26 日	App 数量：10 万余款
			设备激活总量：1050 万

大量智能移动设备接入网络，移动应用爆发性增长使得对数据进行深入挖掘的需求突显，而移动互联网与传统互联网融合，并成为所有媒体的核心节点却是大数据实现的前提。根据 EnfoDesk 易观智库产业数据库最新发布的《2012—2014 中国移动互联网市场预测》数据显示，目前中国移动互联网市场规模已达到 1500 个亿，移动互联网用户超过 5 亿，是 15 年前的 867 倍，互联网普及率达到 39.9%。ZDC 统计数据显示，参与调查者中，使用手机上网者的比例高达 97.4%，仅有 2.6%的调查者表示不使用手机上网。

2. 数据流量剧增，导致网络行业发生新的转变

2013 年 12 月 24 日，据《纽约时报》网站报道，过去一年美国手机产业出现两大趋势：手机网络速度更快，智能手机显示屏更大，其结果是用户的移动数据流量增长近 1 倍。2013 年美国消费者每月使用的移动数据流量由 2012 年的 690MB 增长至 1.2GB；从全球范围来看，消费者每月使用的移动数据流量由 2012 年的 140MB 增长至 240MB。

例如，中国移动数据在 2013 年春节期间涨幅也十分明显，上涨了 105%。据中国移动广东方面透露，总体 GPRS 数据使用量同比增长 63.84%；WLAN 数据量同比增长 227.55%；3G 数据量同比增长 212.68%。

对于如此庞大的数据量，又有哪些是具有商业价值的？怎样挖掘出这些有价值的数据呢？事实上在大数据中，存储在数据库中的结构化数据仅占 10%，邮件、视频、微博、帖子、页面点击等大量非结构化数据占据了另外 90%。怎样从这些与用户行为相关的大数据中挖掘出更多有价值的内容，值得创业者思考和探索，同时也给数据分析与挖掘产业带来更多的机会。

基于如此巨大的数据流量，网站分析（Web Analytics）已成为一种新的火爆产业。Web Analytics 是一种网站访客行为的研究，对于商务应用背景来说，网站分析特指通过来自某网站资料的使用，以决定网站布局是否符合商业目标。例如，哪个登录页面（landing page）比较容易刺激顾客购买欲。这些搜集来的资料几乎总是包括网站流量报告，也可能包括电子邮件回应率、直接邮件活动资料、销售与客户资料、使用者效能资料或者其他自订需求资讯。这些资料通常与关键绩效指标比较，以得到效能资讯，并且还可用来改善网站或者获取营销活动中观众的反应情况。

3. 数据方式在发生转变：数据存储→数据应用

从传统互联网到移动互联网，人们产生的数据越来越多。同时 Google Glass 的诞生让我们有理由相信，未来每个人都将产生更多的数据。但如果仅仅是简单地将这些数

据存储起来，它本身并不具有任何价值。

据统计，目前大数据所形成的市场规模在 51 亿美元左右，而到 2017 年，此数据预计会上涨到 530 亿美元。由此可见，数据背后潜藏着巨大的商业机会。但是，如果大数据时代真的来了，营销人员是否真的能够利用好数据分析，并从中寻找商业价值呢？笔者认为，这是每个企业都应该思考的问题。

4. 互联网营销方式的转变：向个性化时代过渡

正如前面所说，数据结构更加多样化，图像、视频和文档的比例占了半壁江山。大量的用户行为信息记录在大数据中，互联网营销将在行为分析的基础上，向个性化时代过渡。

互联网上，每天新浪微博用户发博量超过 1 亿条，百度大约要处理数十亿次搜索请求，淘宝网站的交易达数千万笔，联通的用户上网记录一天达到 10TB……这些数据运用得好，可以使大众化营销转向个性化营销，从流量购买转向人群购买。

DCCI 提供的数据显示，中国有超过 230 万个网站，网页超 866 亿，移动应用超过 135 万。由此可以预见，国内网络广告投放也将从传统面向群体的营销转向个性化营销，从流量购买转向人群购买。也就是说，未来的市场将更多地以人为中心，主动迎合用户需求。

专家提醒

大数据技术的应用，可以帮助企业从业务的整体设计角度，发展到针对客户的个性化服务，例如，零售企业对于过剩的库存会进行整体促销，如果对于用户购买数据进行分析，就可以针对用户的喜好进行个性化促销，同时也根据用户的购买行为对库存进行准确的调配，以减少浪费。

1.2.5 大数据的发展动力

大数据行业的发展，除了市场需求的驱动和技术水平的进步，还离不开资本与政策的帮助。据麦肯锡报道，大数据已经实现了显著的经济价值：为美国的医疗服务业每年节省 3000 亿美元，为欧洲的公共部门管理每年节省 2500 亿欧元，为全球个人位置数据服务提供商贡献 1000 亿美元，帮助美国零售业净利润增长 60%，帮助制造业在产品开发、组装等环节节省 50%的成本等。大数据体现的巨大经济价值，成功地获得了金融界和政界的青睐。

例如，在英国，虽然经济不景气、财政紧缩，但政府依然为大数据一掷千金。2013 年初，英国商业、创新和技能部宣布将注资 8 亿英镑发展 8 类高新技术，其中 1.89 亿英镑（约 3 亿美元）用于大数据项目。

从目前的实时数据应用状况来看，在许多私企和组织里其实已经开始了大数据应

用，因此这一市场非常需要得到政府的支持。

诸如在线购物等网站已经开始了大数据的应用与实践，例如亚马逊购物网站，系统会根据用户最近的选择和关注过的商品，来进行对应的产品或服务推荐。同理，政府也需要根据这种模式来研究如何将大数据技术应用到公共数据上。

大数据在中国也已驶入“快车道”，政府、企业和科研院所正多方位布局。工信部的物联网“十二五”发展规划，将信息处理技术作为四项关键创新技术工程之一，其中包括海量数据存储、数据挖掘等。随着 4G 牌照在 2013 年末的发放，更高速的网络将带来更大的数据流，为政府和企业带来战略性资源。

例如，国内的政府机构都在推行“智慧城市”这一蓝图。然而，“智慧城市”的信息处理与应用需要具备快速从海量数据中获取决策信息的能力。现代化都市中无所不在的移动设备、RFID、无线传感器以及互联网应用每时每刻都在产生纷繁复杂的巨量数据。

以视频监控为例，一个大型城市目前用于视频监控的摄像头约 50 万个，一个摄像头一个小时的数据量就是几个 G，每天视频采集数据量在 3PB 左右。“智慧城市”的“智慧”主要出自对上述巨量信息的分析、挖掘和处理。大数据技术的应用恰好有效满足了“智慧城市”信息处理需求。如果说具有感知功能的传感器是智慧城市的末梢神经，连接传感器的城市宽带网络是智慧城市的神经系统，那么大数据应用就是智慧城市的大脑，是城市运行的智慧引擎。

综上所述，我们可以看到，大数据成为今天众人瞩目的焦点，是市场、技术、资金以及政府多方因素推动的结果。

1.2.6 展望 2014 的大数据

大数据时代，媒体的转型发展，既是技术问题，也是战略问题，其将对未来的媒体形态和格局产生深远影响。经过 2012 年整整一年的蓄势待发，在 2013 年新年开始时，“大数据”的概念火了，有媒体将 2013 年称为“大数据元年”。

那么，翻过 2013，走进 2014，大数据领域又会向着什么方向发展呢？如表 1-8 所示为 2014 年度大数据发展趋势的预测。

表 1-8 2014 年度大数据发展趋势的预测

发展趋势	具体说明
数据资源化	数据的资源化是指大数据在企业、社会和国家层面成为重要的战略资源。2014 年大数据将成为新的战略制高点，是大家抢夺的新焦点；大数据将不断成为机构的资产，成为提升机构和公司竞争力的有力武器
大数据与云计算的深度融合	大数据处理离不开云计算技术，云计算为大数据提供弹性可扩展的基础设施支撑环境以及数据服务的高效模式，大数据则为云计算提供了新的商业价值，因此从 2013 年开始大数据技术与云计算技术必然进入更完美的结合期

续表

发展趋势	具体说明
Hadoop 将成为企业的关键组件	2014 年，Hadoop 的适用场景将超越批处理和存储，将成为企业数据架构中通用的核心组件，这意味着数据分析将继续成为大数据的首要用例
企业将更加钟情于用户数据	企业将充分利用客户与在线产品或在线服务交互产生的数据，并从中获取价值。为了实现这一点，数据分析能力将比 BI 团队更受重视，它能为企业提供更多的价值
基于海量数据（知识）的智能	2014 年将会有更多基于海量数据（知识）的智能成果出现，甚至有可能产生人工大脑
大数据分析的革命性方法	在大数据分析上，2014 年将出现革命性的新方法。就像计算机和互联网一样，大数据可能是新一波的技术革命。基于大数据的数据挖掘、机器学习和人工智能可能会改变小数据/小世界里的很多算法和基础理论，这方面很可能会产生理论级别的突破
大数据玩转市场决策	大数据将正式登陆市场营销，用于市场营销的大数据技术将在这一年扮演重要角色——影响着广告、产品推销和消费者行为
数据科学兴起	2014 年数据科学作为一个与大数据相关的新兴学科出现，将有专门针对数据科学的专业形成，有博士、硕士甚至本科生出现
数据共享联盟	数据是基础，之前在科技部的支持下，已建立了多个领域的数据共享平台，包括气象、地震、林业、农业、海洋、人口与健康、地球系统科学数据共享平台等。之后，数据共享将扩展到企业层面
大数据新职业	大数据将在 2014 年催生一批新的就业岗位，如数据分析师、数据科学家等。具有丰富经验的数据分析人才成为稀缺资源，数据驱动型工作机将会呈现出爆炸式的增长
云的可视化和控制访问的服务与工具在增多	在 2014 年里，云的可视化将成为安全性的关键。用户希望得到更多的关于云如何运作的可视化信息——无论是基础设施还是 PaaS。现在云仍处于一种“黑匣子”的状态，用户不知道也不理解发生了什么。2014 年将把重点放在使云可视化以及设置访问控制这样的服务或工具上
一个新的分析堆栈将诞生	大数据分析公司 Alteryx 预测，2014 年将出现一个新的数据及分析堆栈，为数据库、分析、可视化提供新的解决方案，这将直接威胁到传统的供应商巨头，而这些供应商也会在匆忙中推出新的解决方案
更大的数据	2014 年，大数据将获得更多的关注、研究、开发和应用，所引起的结果是：体现大数据特征的体量大、速度快、模态多、价值密度低等几个 V 的特性将变得更加极致

1.3 做好准备，大数据面对的挑战

大数据作为一个新生领域，尽管意味着大机遇，拥有巨大的应用价值，但同时也遭

遇工程技术、管理政策、资金投入、人才培养等诸多方面的大挑战。只有解决这些基础性的挑战问题，才能充分利用这个大机遇，让大数据为企业、为社会充分发挥最大价值。

1.3.1 大数据的 12 个不足之处

大数据是信息通信技术发展积累至今，按照自身技术发展逻辑，从提高生产效率向更高级智能阶段的自然生长。无处不在的信息感知和采集终端为我们采集了海量的数据，而以云计算为代表的计算技术的不断进步，为我们提供了强大的计算能力，这就围绕个人以及组织的行为构建起了一个与物质世界相平行的数字世界。

“大数据”术语广泛地出现也使得人们渐渐明白了它的重要性，并渐渐向人们展现了它为学术、工业和政府带来的巨大机遇。大数据时代下的信息技术日渐成熟，但是在高科技发展的今天，也存在着诸多不足，如表 1-9 所示。

表 1-9 大数据的不足之处

不 足 之 处	具 体 表 现
成本问题	数据量的“大”，也可能意味着代价不菲，而对于那些正在使用大数据环境的企业来说，成本控制是关键的问题
带宽能力	运营商带宽能力与对数据洪流的适应能力面临前所未有的挑战
存储技术	大数据处理和分析的能力远远不及理想中水平，数据量的快速增长，对存储技术提出了挑战；同时，需要高速信息传输能力支持，与低密度有价值数据的快速分析、处理能力。硬件的发展最终还是由软件需求推动的，就这个例子来说，我们很明显地看到大数据分析应用需求正在影响着数据存储基础设施的发展
容量问题	海量数据存储系统也一定要有相应等级的扩展能力。与此同时，存储系统的扩展一定要简便，可以通过增加模块或磁盘柜来增加容量，甚至不需要停机
数据平台	部分早期的 Hadoop 项目将面临挑战。有些行业的数据涉及上百个参数，其复杂性不仅体现在数据样本本身，更体现在多源异构、多实体和多空间之间的交互动态性，而当前技术尚难以用传统的方法描述与度量，处理的复杂度很大
延迟问题	“大数据”应用还存在实时性的问题，特别是涉及与网上交易或者金融类相关的应用时。举个例子来说，网络成衣销售行业的在线广告推广服务需要实时地对客户的浏览记录进行分析，并准确地进行广告投放。这就要求存储系统在必须能够支持上述特性的同时保持较高的响应速度，因为响应延迟的结果是系统会推送“过期”的广告内容给客户
个人隐私	大数据环境下通过对用户数据的深度分析，很容易了解用户行为和喜好，乃至企业用户的商业机密，对个人隐私问题必须引起充分重视
商业智能	大数据时代的基本特征，决定其在技术与商业模式上有巨大的创新空间，如何创新已成为大数据时代的一个首要问题

续表

不足之处	具体表现
数据管理	大数据时代对政府制订规则与监管部门发挥作用提出了新的挑战
人工智能	目前，大数据的可视化还没有达到人们的需求
安全问题	某些特殊行业的应用，例如金融数据、医疗信息以及政府情报等都有自己的安全标准和保密性需求。海量数据洪流中，在线对话与在线交易活动日益增加，其安全威胁更为严峻；而且现今黑客的组织能力、作案工具、作案手法及隐蔽程度更上一层楼
人才要求	大数据人才缺乏，大数据时代对数据分析师的要求极高，只有大数据专业化的人才，才具备开发预言分析应用程序模型的技能

除了数据的收集和使用，在大数据时代需要面对的挑战，还有数据的开放。如果说收集数据是一种意识，使用数据是一种文化、一种习惯，那是否开放数据则是一种态度。

1.3.2 大数据挑战的应对策略

当今，大数据的到来，已经成为现实生活中无法逃避的挑战。每当我们要做出决策的时候，大数据就能给我们带来相当大的帮助。但与此同时，大数据也向参与的各方提出了巨大的挑战。对于大数据时代在现如今面临的诸多挑战，笔者也提出几点应对策略，如表 1-10 所示。

表 1-10 大数据挑战的应对策略

应对策略	具体方法
合理获取数据	大数据时代应以智慧创新理念融合大数据与云计算，在大数据洪流中提升知识价值洞察力，实施高效实时个性化运作，建立有效增值的商业模式。另外，还要针对大数据时代的基本特征，加强全方位创新
存储随需而变	与传统的商务智能应用相比，大数据对企业数据的处理能力和商务智能软件提出了更高要求： ➢ 企业必须具备处理大量数据的能力，因为有的企业可能一天之内就要多次处理 PB 级的数据，这是一些传统的存储设备所不能胜任的。 ➢ 传统的数据仓库软件是针对结构化数据设计的，而大数据包含的主要是非结构化的数据，因此传统的数据仓库软件必须改变。 因此，企业可以邀请一些协同处理算法的专家对其用户数据进行分析，从而了解租赁客户的需求
不必急于出台战略性规划和设立产业专项资金	国内的 IT 企业和地方政府已经意识到大数据产业的发展前景，对发展大数据应用有着较大热情。某些城市已经启动了大数据发展战略，计划到 2017 年形成至少 500 亿元的产业规模。在这种情况下，以规划和专项资金等方式进行鼓励，有可能扭曲正常的市场行为，甚至催生泡沫

续表

应对策略	具体方法
筛选与分析大数据	充分利用数据“洞察”自己身边的人或物，在诸多供给方当中精准地匹配自身需求，从而最大限度地满足自身的需求，这样才能真正充分利用大数据实现自身价值的最大化
合理改造、建设和布局 IT 基础设施	对现有的传统数据中心及大量的旧服务器资源，可以通过建立虚拟数据中心或进行就近合并等方式进行改造利用，探索如何通过虚拟化技术和云计算平台管理软件来提高利用效率
培养大数据时代分析的人才	大数据时代对数据分析要求很高，所以培养大数据时代分析的人才势在必行，只有具备大数据专业方面的知识，才能更好地去研究大数据蕴含的特殊技能。人才培养应从高等教育和企业技术人员再培训两个方面入手，允许大学设立大数据相关专业并进行招生，鼓励地方政府出台关于大数据技术人才培训的相关政策
理性面对大数据的价值诱惑	面对社会各界的“大数据”热，应理性分析、冷静观察，扎实做好基础性工作，应充分认识其内在机理及带来的挑战，进一步理清对策思路
云计算和大数据相辅相成	云计算提供计算机资源，如存储、网络容量等，以上所有的能力，使得大数据与云计算相辅相成，成为“最亲密的朋友”
处理好非结构化数据	大数据中，结构化数据只占 15%左右，其余的 85%都是非结构化的数据，它们大量存在于社交网络、互联网和电子商务等领域。由于非结构化数据量猛增，用户必然面临如何同时处理好结构化数据和非结构化数据的问题，例如什么时候将数据放在传统的数据仓库中，什么时候要用开源的 Hadoop 处理数据
提高大数据的可视化	大数据的可视化就是将大数据分析结果转化为公司能够使用的信息。只有大数据分析结果通过可视化处理后，非数据分析专业人士才能够充分理解用语言、图表等表述出来的大数据的信息
安全防范必不可少	通过立法保护个人隐私数据信息应是必由之路。对于公民个人而言，在享受大数据时代所带来的个性化服务的同时，应当加强风险防范意识，在有可能留下隐私数据的情形下要充分考虑由于隐私暴露可能带来的不良后果，并采取相应的防范措施

在大数据时代，数据增长速度加快、数据来源日趋复杂、数据容量迅速扩大、数据类型也变得丰富多样、用户对于数据处理的速度要求越来越高。面对全新的数据业务挑战，企业传统的 IT 建设模式已经无法满足数据增长的需求，因此，新一代数据中心的建设成为未来用户业务发展的根本驱动力。



读书笔记

2

价值：大数据商业变革

学前提示

“除了上帝，其他任何人都应该用数据说话。”不仅是人，整个世界都越来越数据化。信息革命深入发展，如潮的数据澎湃而至，数量之巨，种类之杂，来势之快，前所未有。大数据是推动这场大变革的重要动力，其将成为促进经济社会转型新的关键资源。

要点展示

- 深度挖掘，大数据的商业机遇
- 体现价值，大数据的 4 大变革
- 价值转型，大数据下的商业智能
- 大数据商业变革应用案例

2.1 深度挖掘，大数据的商业机遇

如今，众多企业纷纷进行大数据挖掘，将数据管理变成企业未来 IT 竞争最为核心的力量，而新一代数据中心的建设自然成为 IT 建设的关键。例如，可口可乐公司准备在上海成立一个数据中心，该数据中心主要用于处理中国市场的数据，以此优化企业的业务，并提高行业的竞争力。

可见，在行业互联网化的新 IT 时代，在大数据时代的需求下，数据中心的建设已经成为各行业 IT 建设最为关注的一点，大家都期待借此挖掘大数据的商业机遇。

2.1.1 挖掘大数据的商业价值

通常，企业里面到处都充斥着数据。事实上各行各业的数据量均经历了几何级数的增长，无论是医疗卫生还是金融，抑或是零售业还是制造业。在此类海量数据中，隐藏着无数商业秘密，也孕育着很多机遇以及潜在的成功。

大数据意味着大商机，这是一个大的，可以说是重中之重的事项。对于企业来说，无论是已经开始做大数据了，还是已经开始希望做大数据的项目，研究结果表明：有一个企业或者组织利用大数据技术，另一个企业却没有利用，未来它们的财务状况会出现明显的不同。数据整合带来的价值如图 2-1 所示。

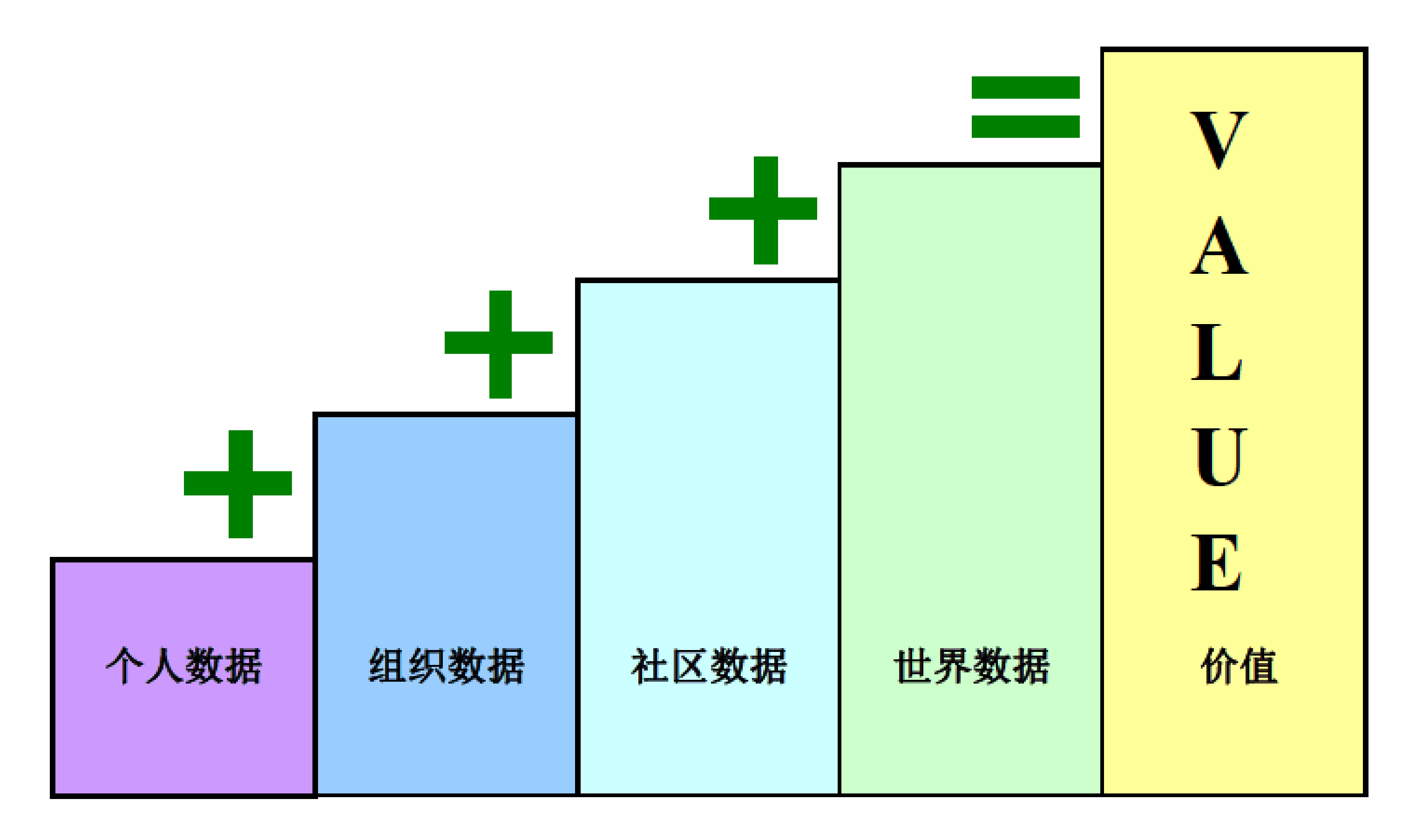


图 2-1　数据整合带来的价值

因此，在今天这样一个数字驱动的大环境下，企业必须能够制定周密计划并且实施可行的解决方案以管理大数据。

当 Twitter 都可以从自己的数据价值中获得不菲的利润，那么任何有大数据的平台都蕴含着极大的商业价值。例如，腾讯 QQ、微信、淘宝、天猫、新浪微博以及视频用户

流量等都是如此。只是企业如何把大数据中的商业价值挖掘出来，并且得以合理地应用却是一个难题，这也是大数据应用的价值所在。可以说，大数据的核心价值理念是商业价值，探求其中蕴含的商业价值对于任何大数据的应用、分析、整合都是非常必要的。

当然，大数据应用和分析最终的目的还是给企业带来更好的收益，技术积累后的优势会在经营中体现出来，这样的结果才是我们需要的。

2.1.2 大数据已进入 4G 时代

如果说 3G 时代，中国追赶世界；那么，4G 来临后，中国正赶超世界。2013 年 12 月 4 日，国家工信部正式向中国电信、中国移动和中国联通发放 4G 牌照，从此开启了中国 4G 网络的商用时代。

很多用户不明白 4G 的概念，下面笔者通过一张图来简单说明一下各种类型网络的区别，如图 2-2 所示。

1G	2G	3G	4G
1G 主要提供语音服务，典型业务就是语音通话	2G 主要提供语音和低速率的数据服务，典型业务是语音通话、短信、彩信等	3G 主要提供语音和相对快速的数据服务，典型业务是语音通话、数据流量	4G 可提供更快速的数据服务，典型业务是高清视频传输、云端游戏等

图 2-2　各种类型网络的区别

伴随着技术的演进，网速得到大幅提升，各种新应用、新服务随之而来。进入 4G 时代之后，移动互联网产业有了更大的想象空间，在突破了“网速”这个瓶颈之后，新型应用的爆发将指日可待。

4G 将使大数据在采集、传输和应用端发生重大变化。信息过载的压力可能会增加，很多数据需要经过处理才能使用，这也催生了大数据产业链上的商机。据了解，4G 最大的数据传输速率超过 100Mbps，是移动电话数据传输速率的 1 万倍。业界认为，4G 将引发一场大数据革命。如图 2-3 所示为 4G 商用对整个通信产业的意义。

4G 时代，大数据的采集和传输速度更快，大数据的体量也会快速膨胀，且会推动大数据存储、计算和分析技术的革新。4G 将使得大数据在采集、传输和应用端都发生非常大的变化，例如，信息过载的压力可能增大，很多数据需要经过处理才能使用，这也催生了大数据产业链上的商机。



图 2-3　4G 商用对整个通信产业的意义

移动网络和大数据是全局零售革命最大的特征。过去的观点是，吸引到店铺来的才是顾客。如今，店铺已经不重要了。由于移动网络的存在，消费者随时可以通过手机或其他移动终端逛商店、下订单或付款，完成购买。

例如，在 2013 年的“双十一”当天，支付宝 350 亿元的成交额刷新了 2012 年的纪录。其中，小米成为最大的赢家，以 5.53 亿元的成交金额位列天猫单店排名第一位，手机销售 33.1 万台，盒子销售 5.6 万台，配件销售 3 553 万元。

小米的胜利不仅仅是其自身营销、价格上的胜利，某种程度上来说，是大时代的胜利。正如小米手机掌门人雷军所说：“小米能成功，首先是因为移动互联网这个大方向选对了。”

专 家 提 醒

笔者认为，在当今时代，物联网担当了数据采集的角色（触角），云存储担当了数据归集和存储的角色（仓库），大数据技术负责收集来的大数据的智能挖掘分析工作（大脑），而互联网技术（包括 4G、光纤等新技术）则是信息传输交换的通道，是信息时代的“高速公路”。

2.1.3 实现商业价值的新捷径

如今，电子商务、社交媒体、移动互联网、物联网的兴起极大地改变了人们生活与工作的方式，它们给世界带来巨大变化的同时，也让一个大数据时代真正地到来。大数据相对于传统数据的优势，主要体现在数据量庞大、数据类型丰富、数据来源广泛 3 个方面，大数据的这 3 大特征不仅仅悄然改变着企业 IT 基础架构，也促使了用户对数据与商业价值之间关系的再思考。

全球知名咨询机构麦肯锡对于不同行业所产生的数据类型进行分析，认为几乎所有行业正在大量产生非结构化数据，如表 2-1 所示。

表 2-1　各大行业的非结构化数据生产频率

行　业	视　频	图　像	音　频	文本/数字
教育	高	中	高	中
政府	高	中	高	高
资源行业	中	中	低	高
建筑	低	高	低	中
公用事业	中	中	低	高
传媒	高	中	高	高
交通运输	中	中	低	高
医疗机构	低	高	低	高
消费休闲	中	低	中	中
专业服务业	中	中	中	高
批发行业	低	低	低	高
零售业	中	低	低	高
流程制作业	中	中	低	高
离散制作业	中	中	低	高
证券投资服务	低	低	低	高
保险行业	低	低	低	高
银行	中	中	中	高
非结构化数据生产频率	高	中	低	

大数据打破了企业传统数据的边界，改变了过去商业智能仅仅依靠企业内部业务数据的局面，其背后蕴含的商业价值不可低估。笔者认为，在大数据时代背景下，企业必须从思维的角度彻底颠覆过去的观点，大数据在未来企业中的角色绝对不是一个支撑

者，而是在企业商业决策和商业价值的决策中扮演着重要的作用。

专家提醒

就像互联网通过给计算机添加通信功能而改变了世界一样，大数据也将改变我们生活中最重要的方面，因为它为我们的生活创造了前所未有的可量化的维度。大数据已经成为了新发明和新服务的源泉，而更多的改变正在蓄势待发。

2.1.4 挖掘大数据的商业机会

随着技术的不断发展，世界已进入大数据时代，而数据背后潜藏着巨大的商业机会。一分钟内，Flicker 上会有 3125 张照片上传，Facebook 上新发布 70 万条信息，YouTube 上有 200 万次观赏。从表 2-1 中可以看出，图片、声音、文字以及这背后用户的习惯和轨迹构成了互联网上的数据资源，大数据时代迎面袭来。

笔者认为，企业要想挖掘大数据的商业机会，一方面，不能将大数据固守在自己的领域里面，要和企业中其他的数据管理、信息分析结合起来；另一方面，在大数据的部署过程中会采用多种技术；最后，大数据需要共同协作和分享来降低成本和风险。

围绕数据的整个产业链上，笔者认为具有以下机会，如表 2-2 所示。

表 2-2 大数据的商业机会

商业机会	具体方案
获得数据	通过把各种行为和状态转变为数据，简称数据化，这是第一个机会，也是基础。大量个人信息数据的获得，这个机会基本属于新浪、微博等这类大企业；大量交易数据的获得，也基本属于京东、淘宝这类互联网企业；小企业基本没机会独立得到这些用户数据
汇集数据	数据的汇集是一个相对复杂的过程，但如果能把各大厂商、微博、政府部门的数据汇集全，这个机会将是极大的
存储数据	汇集了数据后，立即遇到的问题就是存储，这个代价极大，原始数据不能删除，需要保留。因此，提供存储设备的企业，执行存储这个角色的企业，都具有巨大的市场机会，但是这也不属于小企业，或者早期创业者
运算数据	存储完数据后，怎么把数据分发是个大问题，各种 API（Application Programming Interface，应用程序编程接口）、开放平台都可以将这些数据发散出去，用于后续的挖掘和分析工作，这个步骤也需要有大量资本投入，因此不适合小企业
挖掘和分析数据	在转化数据的基础上展开应用，如何把转化数据变为商业机会。需要做增值服务，否则数据就没有价值，因此数据分析和挖掘工作具有巨大的价值，这个机会属于小企业、小团体

续表

商业机会	具体方案
使用和消费数据	电子数据和转化数据的结合应用，在这个含义里面，传统电子数据变为了一种产品，或一种服务。在对数据做到了很好的挖掘和分析后，需要把这些结果应用在一个具体的场合上，来获得回报，做数据挖掘和分析的企业，必须找到这些客户才行，而这些客户肯定也不是小企业

例如，互联网从业者可以运用大数据技术获取和分析用户的消费习惯、兴趣爱好、关系网络以及整个互联网的趋势、潮流。另外，不但社会化媒体基础上的大数据挖掘和分析将会衍生很多应用，而且基于数据分析的营销咨询服务也正在兴起。

专家提醒

不久的将来，数据可能成为最大的交易商品。但数据量大并不能算是大数据，大数据的特征是数据量大、数据种类多、非标准化数据的价值最大化。因此，大数据的价值是通过数据共享、交叉复用获取的。因此，在笔者看来，未来大数据将会如基础设施一样，有数据提供方、管理者、监管者，数据的交叉复用将大数据变成一大产业。

2.1.5 用大数据预测宏观经济

2013 年 5 月，在淘宝网的十周年晚会上，阿里巴巴集团董事局主席马云卸任了阿里集团 CEO 的职位，并做了卸任前的演讲。马云的一番话引起了大家的深思，他说道：“大家还没搞清楚 PC 时代的时候，移动互联网来了，还没搞清楚移动互联网的时候，大数据时代来了。”

从 2009 年 6 月起，和讯网每月推出“和讯预测”系列宏观经济数据，分别邀请十大经济学家和十大券商机构对上月度 CPI、PPI 和当季度 GDP 数据进行预测，并在此基础上建立模型，通过加权平均的方式得出“和讯预测”之“经济学家宏观经济数据预测”和“机构宏观经济数据预测”结果，如图 2-4 所示。

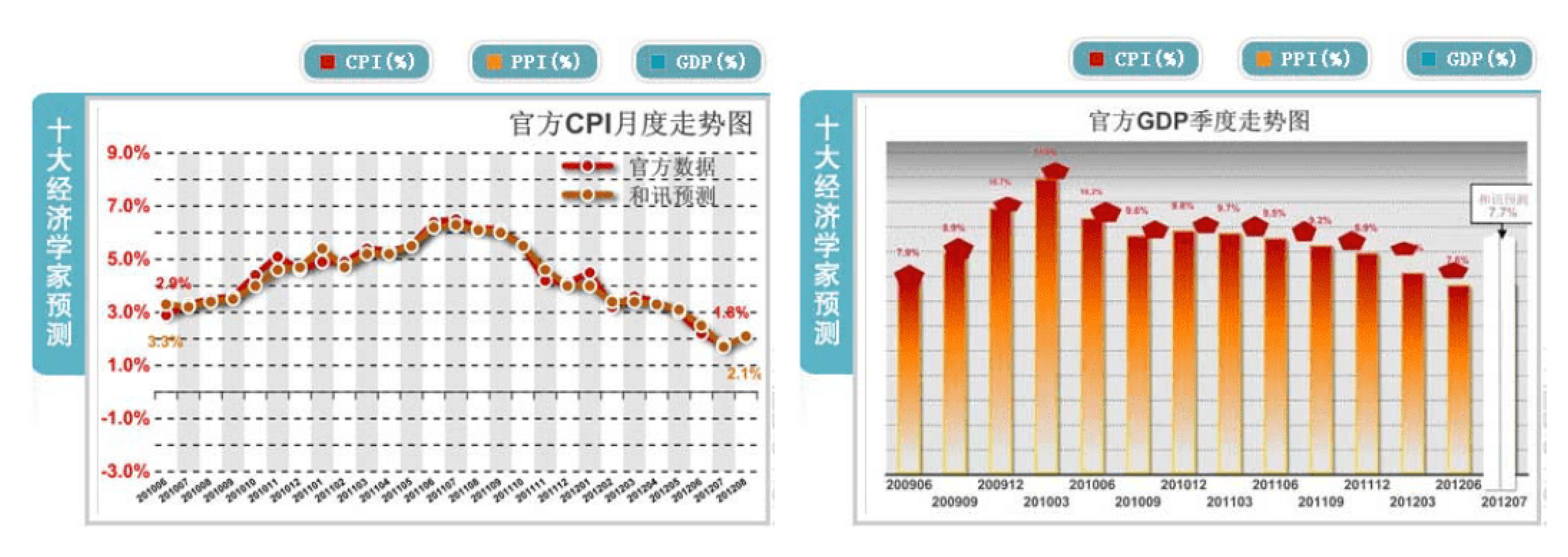


图 2-4　CPI 和 GDP 数据走势图

作为中国财经网络门户，和讯网同时也是政府批准的证券投资咨询机构，并在成立以来的 10 多年里专注收集资本市场与财经领域的信息和数据，因此拥有独立而且丰富的数据库，并且拥有众多学界、机构资源。“和讯预测”的推出既是对这些丰富数据及资源的有效整合，也是希望将这些信息专业加工后可以更好地服务于广大网友，引导投资者理性投资。

2011 年 6 月，东方财富网也推出了“宏观经济数据预测”的业务，汇总十大券商机构对上月度 CPI、PPI、信贷、外贸、工业、投资、消费和当季度 GDP 数据进行预测，并在此基础上建立模型，通过加权平均的方式得出“机构宏观经济数据预测”结果，为网友投资决策提供参考，如图 2-5 所示。



图 2-5　东方财富网上的进口增速走势图

2.1.6　企业用大数据获取优势

如今，数据分析模式正在发生大的转变，当然这一点也为企业带来了真正的机会。大数据平台让所有企业能够通过这种模式转变所提供的洞察力优势，来获得显著的竞争优势。

例如，IBM 在大数据应用和开发方面可以说是处于业界的领先地位。IBM 有 500 多个编程人员和工程师，以及 15000 次的 IBM 客户参与，而且 IBM Power Systems 全线产品均可运行 Linux。作为 IBM Power Systems 旗下的一条子产品线，Power/Linux 可以通过更少的处理器数量提供更好的系统性能，满足大数据、开源和行业解决方案工作负载的需求，帮助企业尽展大数据分析洞察智慧。

也许你还没有看到大数据到底有何优势，那么下面再举一个典型的案例。作为全球知名硬件产品、解决方案、云计算服务的提供商——中科曙光，推出了曙光行业大数据

系统，这是一个能够感知和度量数据的、全面互联互通的系统，其能够快速、智能地分析海量数据，以提高洞察力并帮助企业做出明智决策，为客户提供创新的产品和服务，如图 2-6 所示。

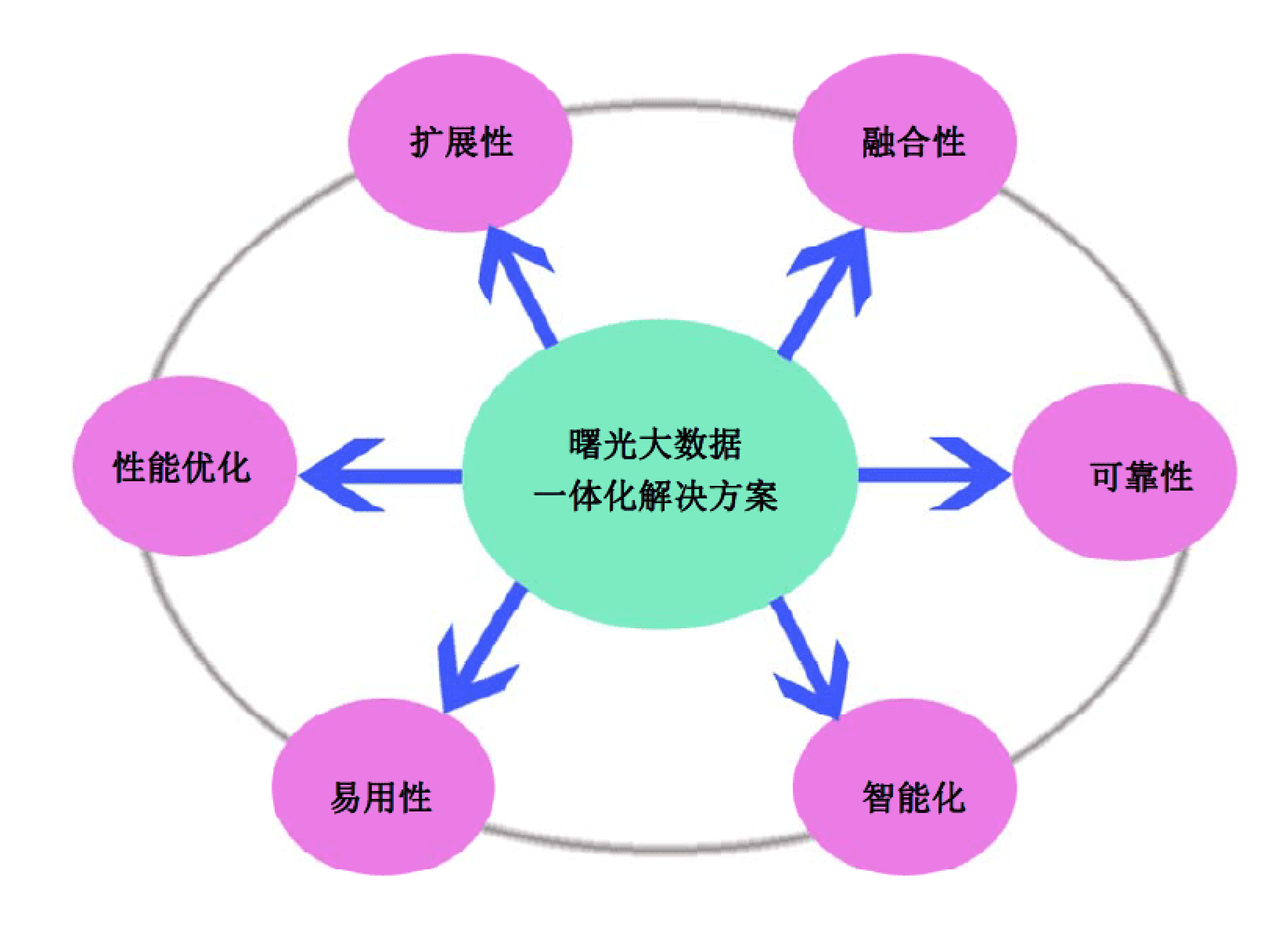


图 2-6　曙光行业大数据系统的竞争优势

2.1.7　大数据有待更深的挖掘

大数据并不是新的概念，在移动互联网发展起来后，数据增长速度加快，整个产业压力突出，传统数据库技术已无法满足运营商对大数据充分利用的需求，在此背景下，大数据成为近年来的热点。

大数据时代主要是对技术的综合运用和对数据的深度挖掘。尤其是对于运营商来说，大数据带来的机会大于挑战。运营商有自己的网络，积累了大量非常有价值的数据，可以进行客户分析。利用网络收集数据，对运营商运营方式的改变是个机会。

例如，电信运营商不仅可以利用自身在运营网络平台的优势，更可以突破传统模式，发展大数据分析服务、移动营销等高端大数据业务。随着大数据的技术成熟和应用的推广，运营商将可以围绕数据标准化、精准营销、优化用户服务体验、提高业务效率等 4 个方面来强化大数据的应用，提高运营商在企业和个人用户中的影响力，如图 2-7 所示。

专 家 提 醒

大数据的应用可以帮助人们不再追求精妙的算法，而是以过去所有的数据为基础来准确推断和判断未来可能发生的事情。因此，企业如果能够通过技术的进步，不断释放大数据的潜在力量，其将会成为未来数字时代中最大的赢家。

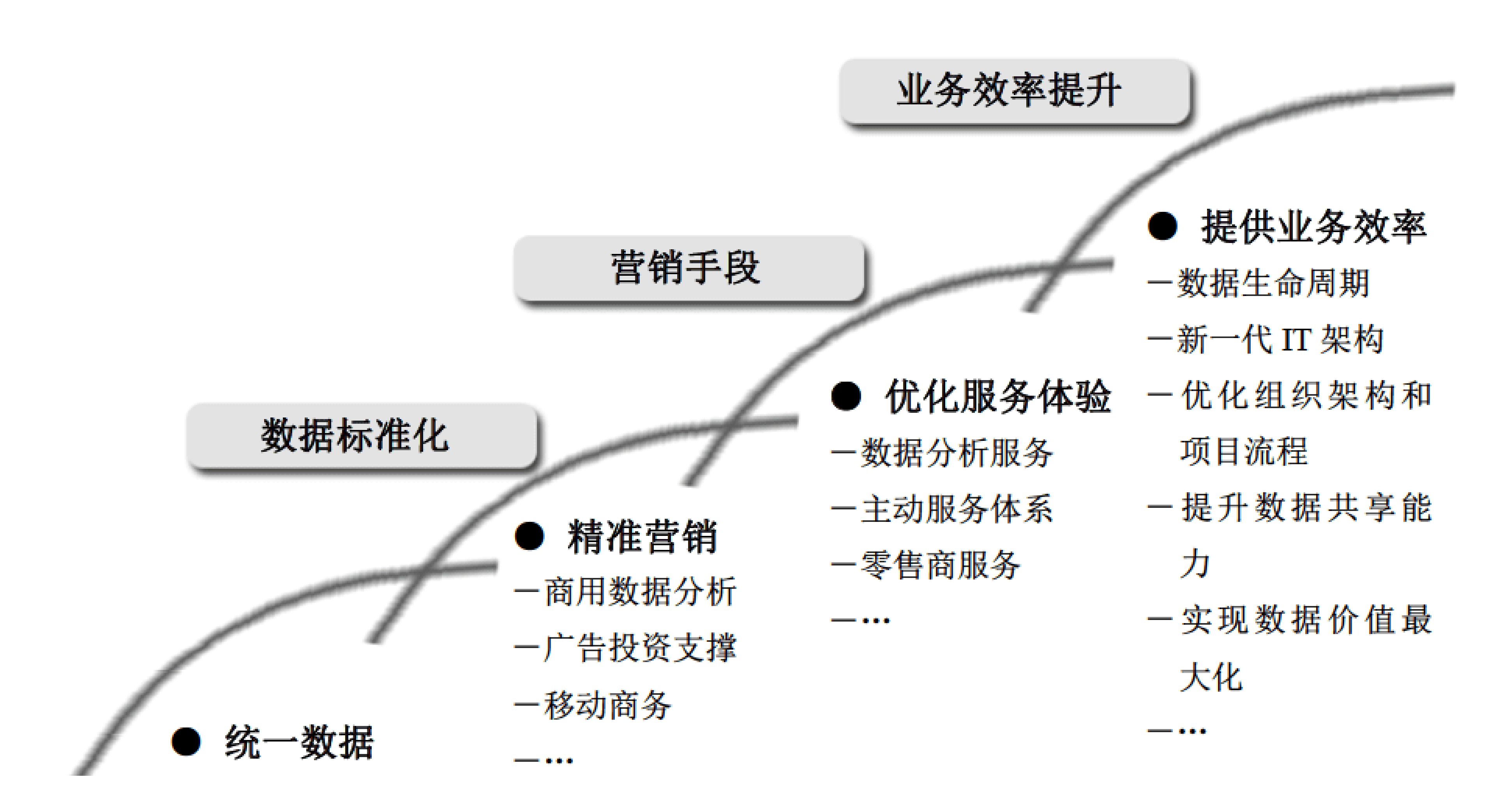


图 2-7　电信运营商可以更深层次地挖掘大数据的价值

2.2　体现价值，大数据的 4 大变革

大数据即将开创信息社会的崭新时代，它能够改变我们看待世界的方式。那么大数据意味着什么，它到底会改变什么？笔者认为，仅从技术和商业的角度回答，已不足以解惑。大数据只是宾语，离开了人这个主语，它再大也没有意义。因此，我们需要把大数据放在人的背景中加以透视，理解它作为时代变革力量的所以然。

2.2.1　变革医疗卫生

大数据的影响也已经渗透到各个行业的应用当中，最具代表性的行业有互联网、电商、金融、公共服务等，当然其中也包括医疗服务。

医疗卫生行业作为典型的传统行业，其 IT 网络的建设具有一定的行业复杂性与特殊性。但是，随着医疗改革的逐步深入，医疗服务质量的提高相比于医疗服务效率的提升更加重要。那么，如何在众多医疗机构中突出自己的特色，做到真正的急患者所需，更好地为患者服务，才是医院管理层真正关注的关键。

在过去的 30 年间，我国的医疗行业经历了医改、新医改，医疗信息化也经历了数字化、“四梁八柱”、35212 工程等不同的发展阶段，信息技术的发展使数字化医疗日趋成熟。云计算、大数据等新兴技术的推动又给医疗信息化和新医改带来了新的契机。

专 家 提 醒

“四梁”是指 4 大医药卫生体系：全面加强公共卫生服务体系建设；进一步完善医疗服务体系；加快建设医疗保障体系；建立健全药品供应保障体系。“八柱”是指以下 8 大医疗

卫生改革：建立协调统一的医药卫生管理体制；建立高效规范的医药卫生机构运行机制；建立政府主导的多元卫生投入机制；建立科学合理的医药价格形成机制；建立严格有效的医药卫生监管体制；建立可持续发展的医药卫生科技创新机制和人才保障机制；建立实用共享的医药卫生信息系统；建立健全医药卫生法律制度。

例如，一个普通的三甲医院每天就要接待上万名患者，患者的基本信息、影像信息与其他特殊诊疗信息汇集在一起就形成了一个庞大的数据库。日积月累，这个数据量将会以几何数字倍增，为医院的数据存储、集成、调用等应用都带来了巨大压力。因此，怎么才能精确管理与快速调用这些数据为医生和管理层所用，成为了目前很多医院 CIO 都关注的热点。

大数据的到来，使很多医院高管们不再靠差不多、经验和直觉习惯做决策，逐步转变思维方式，通过对海量数据的挖掘和运用，更多地基于事实与数据分析做出决策。这对信息技术人员来说是机遇也是挑战，而这些影响都是大数据带来的。

2.2.2 带来商业革命

大数据不仅改变了医疗卫生领域，整个商业领域都因为大数据而重新洗牌。

在此，笔者首先要告诉大家一个“启动内需”的原理：生产者是具有价值的人，而消费者是生产者价值的意义所在。有意义的才有价值，消费者不认同的，就卖不出去，就实现不了价值；只有消费者认同的，才卖得出去，才实现得了价值。然而，大数据可以帮助我们从消费者这个源头识别意义，从而帮助生产者实现价值。

例如，华声财讯信息技术有限公司结合云计算、大数据时代的发展趋势，推出了基于 SMAS（社会化媒体云服务平台）的新一代“企业舆情监测系统”，为客户量身打造全媒体时代的防御利器，把握数据挖掘和业务情报产业的先机，如图 2-8 所示。



图 2-8　华声财讯的大数据舆情监测业务

2.2.3 改变人们思维

中国科学院的怀进鹏院士在“第五届中国云计算大会”发表了题为“大数据与大数据的科学与技术问题”的主题演讲，他在演讲中表示：“大数据的发展可能会改变经济和社会生活，可能会改变科学研究的途径，甚而改变人类的思维方式。”

互联网重塑了人类交流的方式，而大数据则不同，它标志着社会处理信息方式的变化。随着时间的推移，大数据可能真的会改变我们思考世界的方式。随着我们利用越来越多的数据来理解事情和作出决定，我们很可能会发现生活的许多层面是随机的，而不是确定的。

专家提醒

大数据的确改变了我们的思维，更多的商业和社会决策能够“以数据说话”。不过抛开这所有的利好，如何让大数据不侵入我们的隐私世界，也是与之伴生并需严肃考虑的问题。

2.2.4 开启时代转型

大数据的核心就是预测，相关关系可以帮助我们捕捉现在和预测未来，其带来的技术变革将开启一次重大时代转型。

例如，百度搜索指数显示，自 2013 年 6 月至 9 月，“考研”相关搜索词累计达到了 1.15 亿，日均接近 100 万，较 2012 年同期增长 10%，如图 2-9 所示。依据 2012 年 176 万的考研报考人数，百度大指数预测，2013 年考研的报考人数较 2012 年相比还会增长，预计能突破 190 万。

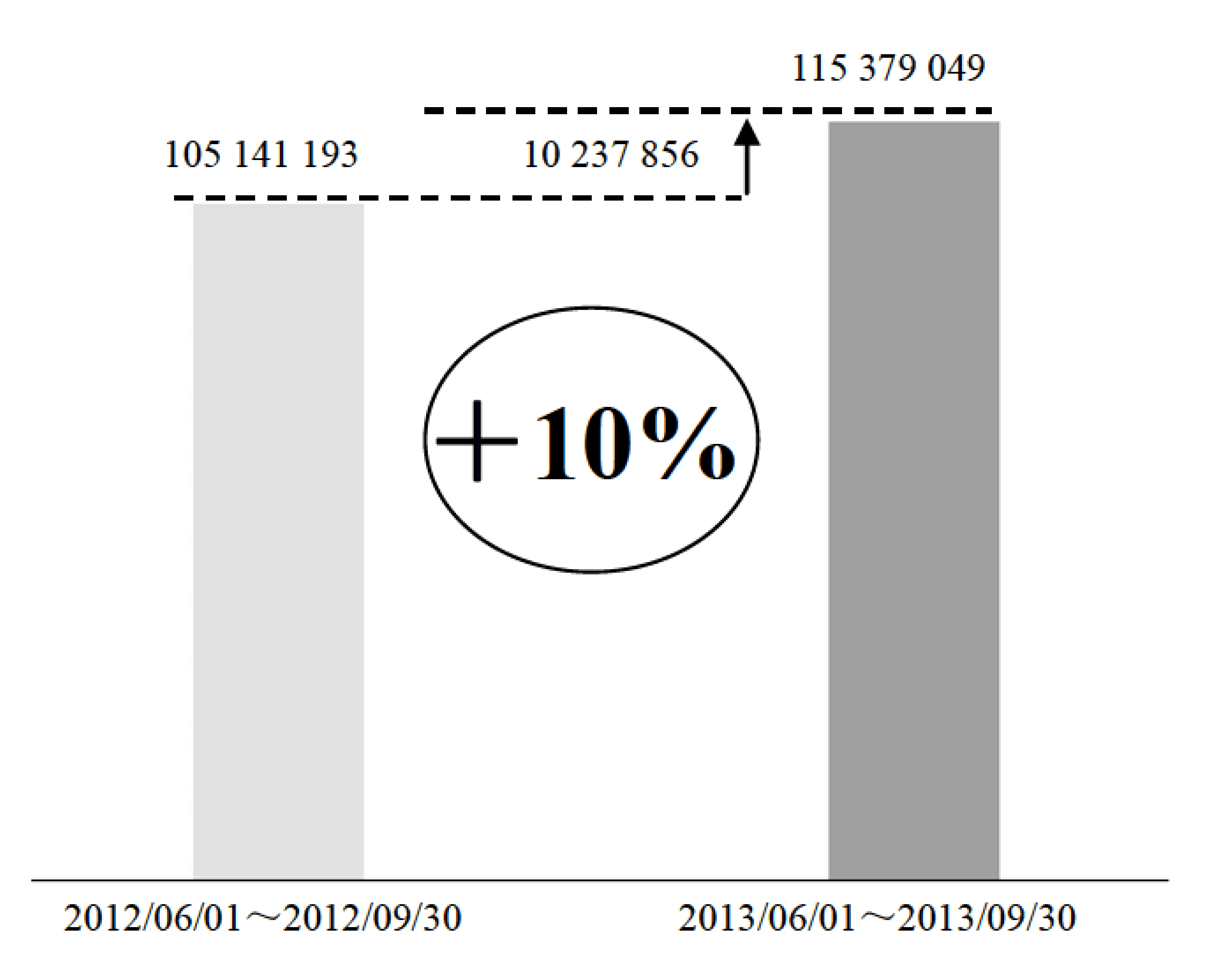


图 2-9　2013 年百度考研搜索指数较 2012 年同期增长 10%

A 和 B 事件如果经常在一起发生，那么注意到 B 发生，就能预测 A 也发生。这种关系已在零售业和电子商务中被广泛运用。例如，某家便利店通过分析零售终端的数据，得出了“温度低于 15 摄氏度暖宝宝的销售量便增加 5%”的相关关系。于是，只要温度低于这一度数，店内的暖宝宝就会上架。

专家提醒

大数据时代最大的转变就是，放弃对因果关系的渴求，取而代之关注相关关系。也就是说只要知道“是什么”，而不需要知道“为什么”。这颠覆了千百年来人类的思维惯例，对人类的认知和与世界交流的方式提出了全新的挑战。

2.3 价值转型，大数据下的商业智能

如今，也许你并不了解大数据，但大数据的应用确实已经遍地开花。例如，金融行业通过大数据来鉴别个人的信用风险；快递领域通过数据来确定行驶路线，减少等候时间；政府通过大数据来找出最容易发生火灾和井盖爆炸的地点；商场通过大数据发现产品之间的关联。在大数据时代，一切都存在着可能，智能商业带来的价值转型正在悄然发生，而我们也正在体验这一切改变。

2.3.1 大数据为商业智能构建基础

DBA（Database Administrator，数据库管理员）们都知道数据在任何商业智能（Business Intelligence，BI）解决方案中都是最重要的部分。

商业智能作为一个工具，是用来处理企业中现有数据，并将其转换成知识、分析和结论，帮助业务或者决策者做出正确且明智的决定的。商业智能是帮助企业更好地利用数据提高决策质量的技术，其包含了从数据仓库到分析型系统等。

大数据 BI 是能够处理和分析大数据的 BI 软件，区别于传统 BI 软件，大数据 BI 可以完成对 TB 级别数据的实时分析。例如，阿里巴巴敏锐地捕捉到大数据的巨大潜能。2012 年，阿里巴巴提出大数据战略，通过资源共享与数据互通创造商业价值。在 2012 年的“双十一”销售热潮中，阿里巴巴以云计算为基础的数据服务，对数以亿万计的消费者需求信息进行捕捉，帮助网商随时调整销售决策。

如今，新一代信息技术已经彻底地改变了 BI 市场环境，微博、云计算、物联网、移动互联网等各种爆炸式数据，给商业智能的蓬勃发展提供了良好的“大数据”基础。

大数据为 BI 带来了海量数据。对挖掘来说，大数据量更容易对比，它加速了 BI 效率和整合能力的提升。因此，有人大胆预测：与大数据相关的商务智能分析将引领管理信息化的发展。

2.3.2 Oracle BIEE 商业智能系统

Oracle BIEE 是 Oracle 商业智能平台企业版，由收购、整合 SIEBEL 和 HYPERION 相关 BI 部分组建形成，在 Oracle 整个商业智能体系架构中主要承担数据分析应用和可视化展示工作。Oracle BIEE 架构如图 2-10 所示，其中最重要、最核心的是 BI Server 和 BI Server 所操作的 Repository。

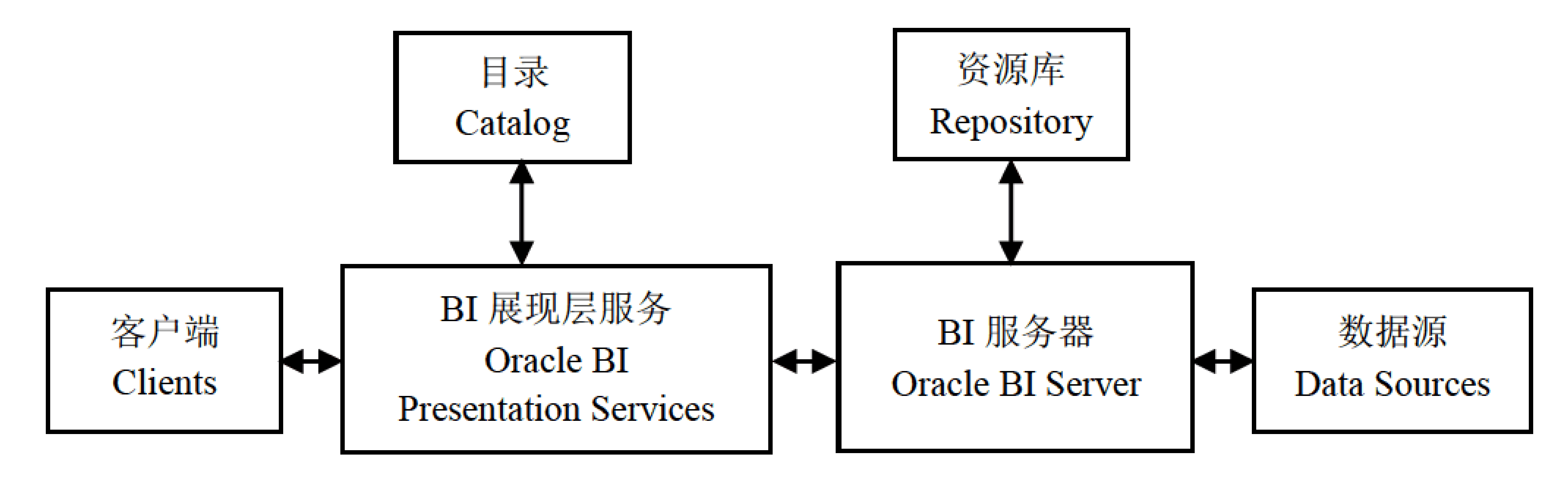


图 2-10 Oracle BIEE 架构

利用 Oracle BIEE 可以将商业智能分析模型清楚简洁地展现出来，开发人员在定义好元数据后，业务人员即使了解内部库表和相关技术，也可以以一种可视化的、简单的方式产生出自己所需要的智能数据报表，这大大提高了经营分析的效率，如图 2-11 所示。同时，随着云计算技术的不断发展，给商业智能行业带来了新的启示。基于云计算的商业智能平台可以作为 Web 服务提供给用户，商业智能的 Web 化和服务化，或将成为一个新的趋势。

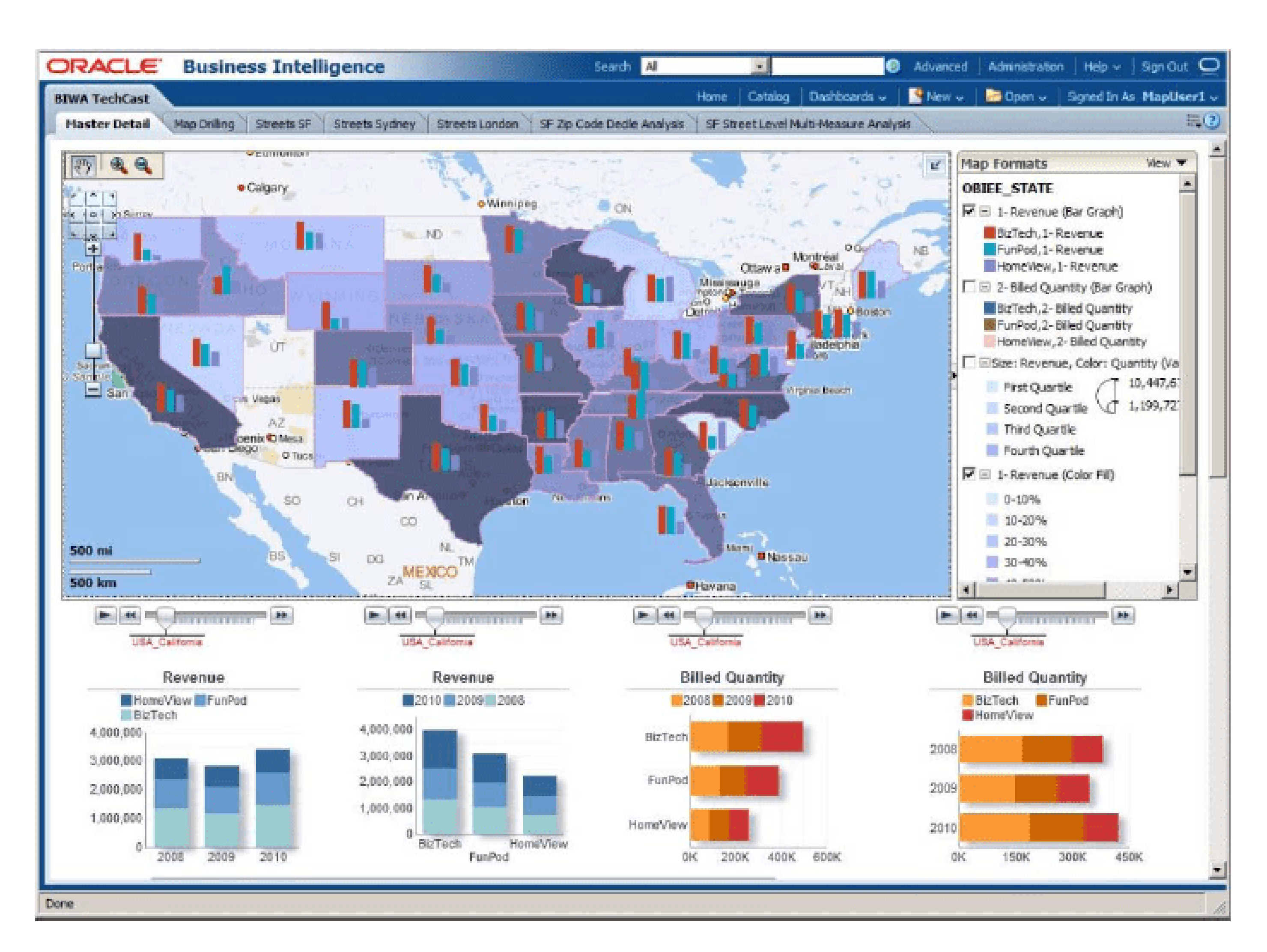


图 2-11 基于 Oracle BIEE 的商业智能分析系统

2.3.3 商业智能成就行业价值机会

1989年，商务智能界“教父”——Howard Dresner提出“商业智能”的概念，不久后便被人们广泛了解。当时将商业智能定义为一类由数据仓库（或数据集市）、查询报表、数据分析、数据挖掘、数据备份和恢复等部分组成的，以帮助企业决策为目的的技术及应用。

在大数据时代，企业如果想要抢夺大数据市场，就需要具备一定的实力，然而报表的呈现和简易分析只是停留在“B”的阶段，要想达到“I”的阶段，必须要结合整个大环境、大行业的数据来判断分析并给出真正有价值的信息和决策建议，这取决于你能拿到多广多深的数据和你的数据挖掘分析能力以及建模能力。

商业智能与大数据的区别在于，大数据能够基于BI工具进行大容量数据处理和非结构化数据处理，与传统基于事务的数据仓库系统相比较，大数据分析不仅关注结构化的历史数据，它们更倾向于对Web、社交网络、RFID传感器等非结构化海量数据进行分析，大数据无疑是对BI的一个完美补充。

例如，2002年，民航旅客量突破一个亿，这一个亿旅客带来了海量数据的增长，而且数据类型也是丰富多样，所以在那个时候，航信团队就认为数据挖掘是非常必要的工作，利用数据仓库平台做了早期的挖掘。之后经过调研，IT团队也采用了专业商业软件去部署，这个平台也给客户带来了很多价值。

商业智能通常被一些大企业作为强有力的掘金石，在实现信息化建设后，进而贯彻决策的解决方案，而在当前中小企业应用的商业智能的过程中还存有一定的瓶颈，中小企业的实施成本及对商业智能的认识及发展力度还存在一定差异。

据Gartner（全球最具权威的IT研究与顾问咨询公司，成立于1979年）透露，BI市场正在以每年9%的速度增长，到2014年市场价值将高达810亿美元，2020年将增长至1360亿美元。

专家提醒

企业信息化已逐渐由传统运营层管理转向决策层管理，企业实施BI犹如试穿鞋子，企业BI应用的核心取决于企业决策与业务优化，企业对于BI的深化，需要具备一定的信息化基础，BI应用是基于业务优化、运营管理与决策的基础上的。

2.3.4 BI导出商业潜能和社会走向

如今，传统数据仓库的性能已无法应付庞大的信息，但是大数据技术使我们能够访问和使用这些宝贵的、大规模数据集，以应对越来越复杂的数据分析和更好的商业决策。

例如，当你在听音乐时，豆瓣电台会推荐你可能喜欢的音乐；当你在当当网下单某

本书时，它会提醒购买这本书的人中有30%也购买了另外一本书（如图2-12所示），这些都是基于大数据分析的。大数据带来的另一改变是，更多事物可以数据化。购物习惯可以数据化，社交关系可以数据化，社会热点的走向也可以数据化（通过对搜索关键词的分析）。这些数据可以导出商业潜能，更能导出社会走向。



图2-12　当当网的购书提醒功能

随着互联网技术的发展，未来的大数据时代，将是各种信息呈现规模化快速增长的状态。如何更快获取有用的信息是关键，智能分析工具会变得越来越重要，其可以凌驾于多个管理系统、数据库之上。如何通过更灵活、可控的BI工具，真正挖掘出大数据时代的价值，是大数据和BI面临的共同挑战。

2.3.5　商业智能的6大发展前景

总体上来看，商业智能的发展有以下几个特点：实时、操作型、与业务流程的集成、主动以及跨越企业边界等。商业智能的实时特性，可以让公司与顾客拉近距离，而实时商业智能可以迅速地处理数据，并给出及时、有效的决策。

如今，商业智能的概念从技术到应用都发生了巨大的变化，从商业智能到商业分析，再到企业绩效管理，然后再到企业绩效优化。那么商业智能的发展在技术上和应用上的趋势如何呢？笔者在这里谈谈自己的观点，如表2-3所示。

表 2-3　商业智能的发展前景

发 展 前 景	趋 势 预 测
内存分析	内存技术已经成为了万众瞩目的焦点，它能够为不断增长的庞大数据提供快速分析。未来，大型企业会逐渐采用如 HANA 及 Exalytics 之类的高端应用，然而大多数客户会继续采用 QlikTech、Microsoft（Power Pivot）及 Tableau 等供应商提供的灵活的内存解决方案，或如 MicroStrategy 及 IBM Cognos 使用方法之类的纯软件解决方案
可视化发现	可视化发现技术会成为商业智能的重头戏。可视化发现不同于内存技术，尽管在有些行业将两者混同，而且不少可视化发现工具也内置了内存引擎
大数据	大数据会导致硬盘读取数据非常慢，所以大数据需要一个快到秒级的、让用户感觉无缝对接的平台，并且还要让业务人员尽可能通过简单方式来使用这个平台。大数据让更灵活的框架和拥有灵活数据挖掘算法的商业智能解决方案，拥有了更广阔的发展空间
移动 BI	移动 BI 性能将继续提升，更多 BI 供应商将调整应用，以适应移动 BI。例如，平板电脑能够支持线下或飞行模式，提供更高的安全性以及更好的性能
云计算 BI	不少供应商将云计算视作减少内存消耗的最佳方法，称其能够在计算高峰时期提供灵活的数据解决方案
协作型商务智能	从数据出发，可以在供应商、企业内部和客户之间共享分析的结果，通过结果发现某些行动可能产生的风险，这些风险会给供应商、企业内部、客户带来损失

2.4　大数据商业变革应用案例

人们懵懂地意识到，数据即将成为改变未来社会的重要力量。然而，大数据究竟改变了什么，在人们脑中仍是个模糊的影子。那么，通过本节的应用案例，笔者来告诉大家大数据到底带来了什么样的商业变革。

2.4.1　【案例】大数据助力地产行业

中国建筑第五工程局有限公司（以下简称中建五局），不但是世界 500 强企业，也是中国最具国际竞争力的建筑地产集团——中国建筑工程总公司的成员企业。

由于中建五局现有的 ERP 系统不能将原始数据进行加工，给管理者提供有价值的辅助决策信息，也不能以更加丰富的形式展现运营状态，因此，中建五局准备在全局范围内搭建一套企业经营决策分析系统。2013 年 7 月 9 日，“中建五局管理信息化集成系

统”项目验收会在长沙举行，经过验收委员会专家评审，由用友软件与中建五局合作开发的中建五局管理信息化集成应用系统顺利通过验收。

用友软件通过对全局的战略、经营、财务、项目运营以及风险预警等分析体系的建立，为中建五局提供多种关键指标对比、趋势分析，并能够从不同的维度对数据进行统计分析，挖掘数据信息，为企业提供决策支持依据，如表 2-4 所示。

表 2-4　中建五局管理信息化集成系统的基本功能

基 本 功 能	具 体 内 容
战略分析	从全局角度出发，对包括房屋建筑、基础设施、房地产开发等多个业务板块，从战略市场、重要市场以及海外市场进行各种经营指标的同比、环比以及对比分析，以及同其他兄弟单位的对比体系分析
经营分析	从下属单位考核机制来着手，对其二级分支机构、三级分支机构以及项目部三个层面进行包括利润、费用、成本、收益等多个方面的数据分析，并且也从区域、时间、专业、行业等多种维度去分析数据
多种关键指标对比	根据关键指标，主要从纵向、横向的角度对财务数据进行分析。纵向是指从历史的角度进行分析，横向是指从行业角度分析和集团内部中各个分公司的对比分析
竞争力分析	建立在企业核心资源基础上，包括对人力、技术、装备等多方面因素的反映，重点反映的是 HR 情况
项目运营分析	是中建五局的经营决策分析系统中的一个核心模块，主要体现在项目实施阶段中对项目的整体把握，使高层管理人员能够对项目的运营情况有一个直观了解，并且以此为依据做出正确的决策
风险预警	对关键的指标进行必要的预警提示。风险预警统一采用时间单位，随着时间推进，预警值发生变化，达到预设值标线则进行预警

【案例解析】：在本案例中，中建五局管理信息化集成系统涵盖了大型建筑企业集团的主要管理内容，建立了从上到下的主数据标准化体系和基于 ESB（Enterprise Service Bus，企业服务总线）的便于扩展数据的交换体系，将不同运行系统的服务通过定义好的接口联系在一起，实现不同业务以一种统一和通用的方式进行自由交互。

2.4.2　【案例】大数据预测机票价格

美国工程师奥伦·埃齐奥尼（Oren Etzioni）搭飞机时，发现旁边的旅客买票比他便宜。于是埃齐奥尼开发了一个 Farecast 工具，用于预测机票价格的波动。

通过预测机票价格的走势以及增降幅度，Farecast 票价预测工具能帮助消费者抓住最佳购买时机。由于 Farecast 的运转需要海量数据的支持，埃齐奥尼找到了一个行业机

票预订数据库。依靠这个数据库进行预测时，预测的结果是基于美国商业航空产业中，每一条航线上每一架飞机内的每一个座位一年内的综合票价记录而得出的。如今，Farecast 已经拥有约 2000 亿条飞行数据记录。

截至 2012 年，他的 Farecast 系统已经可以用网上的 10 万亿条价格记录去推测机票何时价格为何，预测准确度达 75%，帮助旅客平均每张机票节省 50 美元。

Farecast 是大数据公司的一个缩影，也代表了当今世界发展的趋势。五年或者十年之前，奥伦·埃齐奥尼是无法成立这样的公司的。他说："这是不可能的。"那时候他所需要的计算机处理能力和存储能力太昂贵了！虽说技术上的突破是这一切得以发生的主要原因，但也有一些细微而重要的改变正在发生，特别是人们关于如何使用数据的理念。

【案例解析】：如今，人们已不再认为数据是静止和陈旧的。但是在以前，一旦完成了收集数据的目的之后，数据就会被认为已经没有用处了。比方说，在飞机起飞之后，票价数据就没有用了。

现代商业环境变幻莫测，因此，对于企业来说，在大数据时代做好准备，利用好大数据尤为重要。

2.4.3 【案例】用大数据增强竞争力

2002 年，北京移动开始构建 IDC（Internet Data Center，即互联网数据中心）。经过此后 8 年的努力，一共建设了 8 个重要 IDC 核心节点，机房建设面积一共是 4 万平方米，有上百 G 的带宽连到骨干网上。

北京移动拥有比较丰富的 IDC 运营经验和实力，是国内首家通过 ISO 27001 认证的数据中心，早在 2003 年的时候已经通过 BS79 认证，在 2004 年年底的时候申请到 ISO 27001 这样的认证标准。北京移动 IDC 也是中国移动最主要的内容枢纽中心之一，担负着疏通全网内容的重要战略使命，现在整个中国移动 6 个亿的 Web 访问前十名站点都在北京移动数据中心之内。

在大数据领域，北京移动的标杆企业是云基地。在 BI 系统的支持下，北京移动逐步强化"用数据说话"的工作理念，巩固了业务运营的数据支撑优势，增强了企业的核心竞争力。BI 系统就是北京移动打造的另一只金翅膀，助力企业展翅高飞。

北京移动 BI 系统的成功在于以下两个方面：

- 帮助业务部门建立了数据分析和精细化应用的框架体系，从企业全局来支撑日常的数据分析需求。
- 以高端客户服务为起点，建立一系列 BI 专题来促进高端客户服务更加精细化、个性化、人性化，推动了高端客户服务模式变革，逐步建立起以"客户为中心"的跨部门协作的服务体系。

总之，BI 系统的应用提升了企业的运营效率，保障了业务高效地开展，如图 2-13 所示。

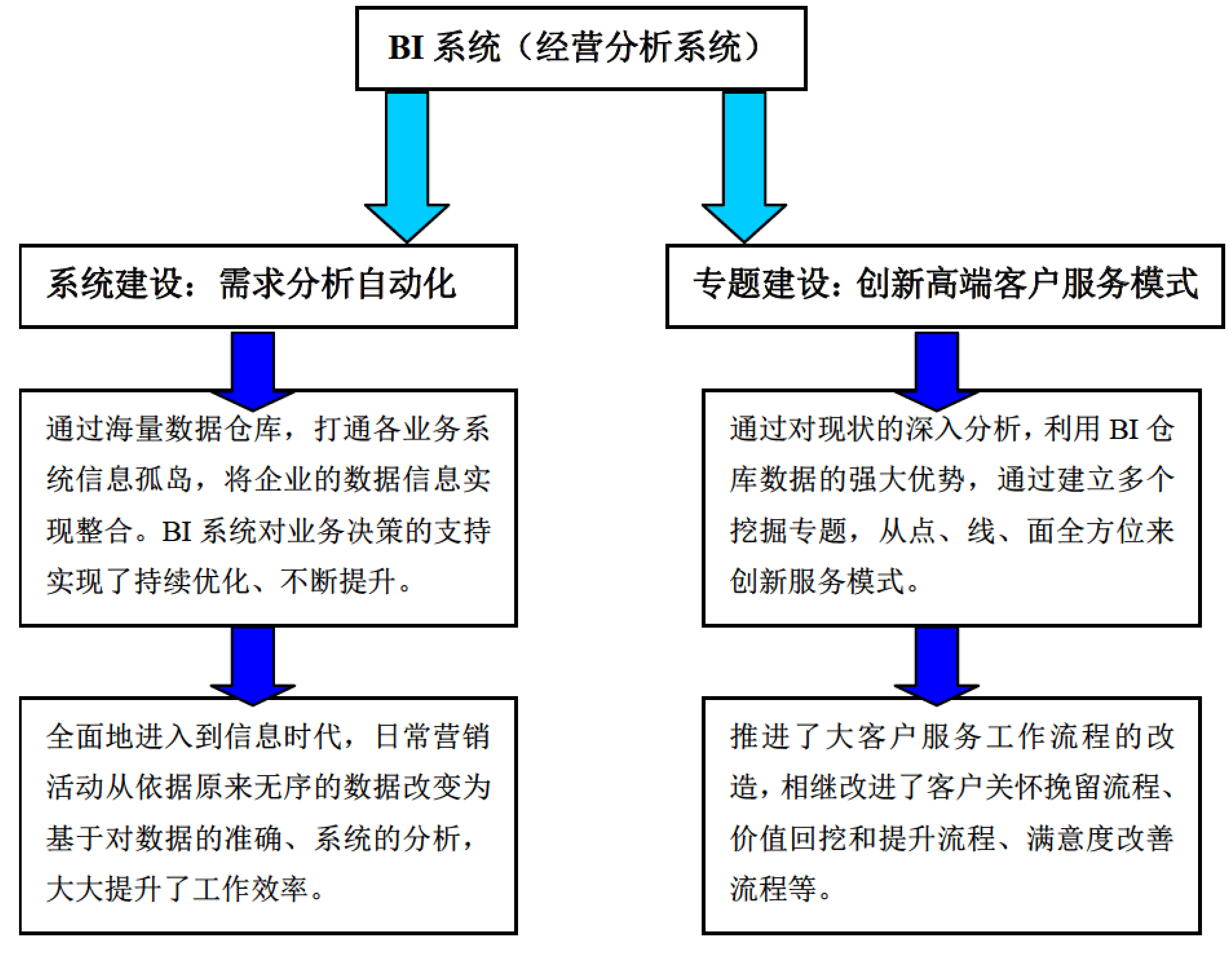


图 2-13　BI 系统的应用

【案例解析】：移动互联和大数据时代的到来，极大地改变了企业传统的经营模式、经营环境和经营方式，如何抓住新的机遇、应对新的挑战成为企业必须面对的问题。在本案例中，北京移动在移动互联和大数据商业环境下，利用商业智能的优势，来扩大市场、降低成本、提升效率、应对危机、获得机遇，并实现跨越式发展。

2.4.4　【案例】大数据助力企业管理

上海帝高绒毛服饰有限公司（简称帝高羊绒）创立于 1989 年 1 月，其凭借精湛的工艺技术和先进的管理经验，经过几年的发展造就了享负盛名的“百纯帝高”羊绒衫。2003 年 10 月，帝高羊绒开始使用博科商业智能——财务智能仓系统（BI-FIW），希望通过商业智能来建立起企业历史管理数据之间的相互关系，满足企业快速决策的管理需要，如图 2-14 所示。通过 3 年的逐步建设，帝高羊绒信息化数据已经涉及采购、销售、库存、往来、总账等业务内容。在此过程中，博科资讯的实施人员对帝高羊绒的数据仓库进行了进一步升级，以满足商业智能系统的运行需要。

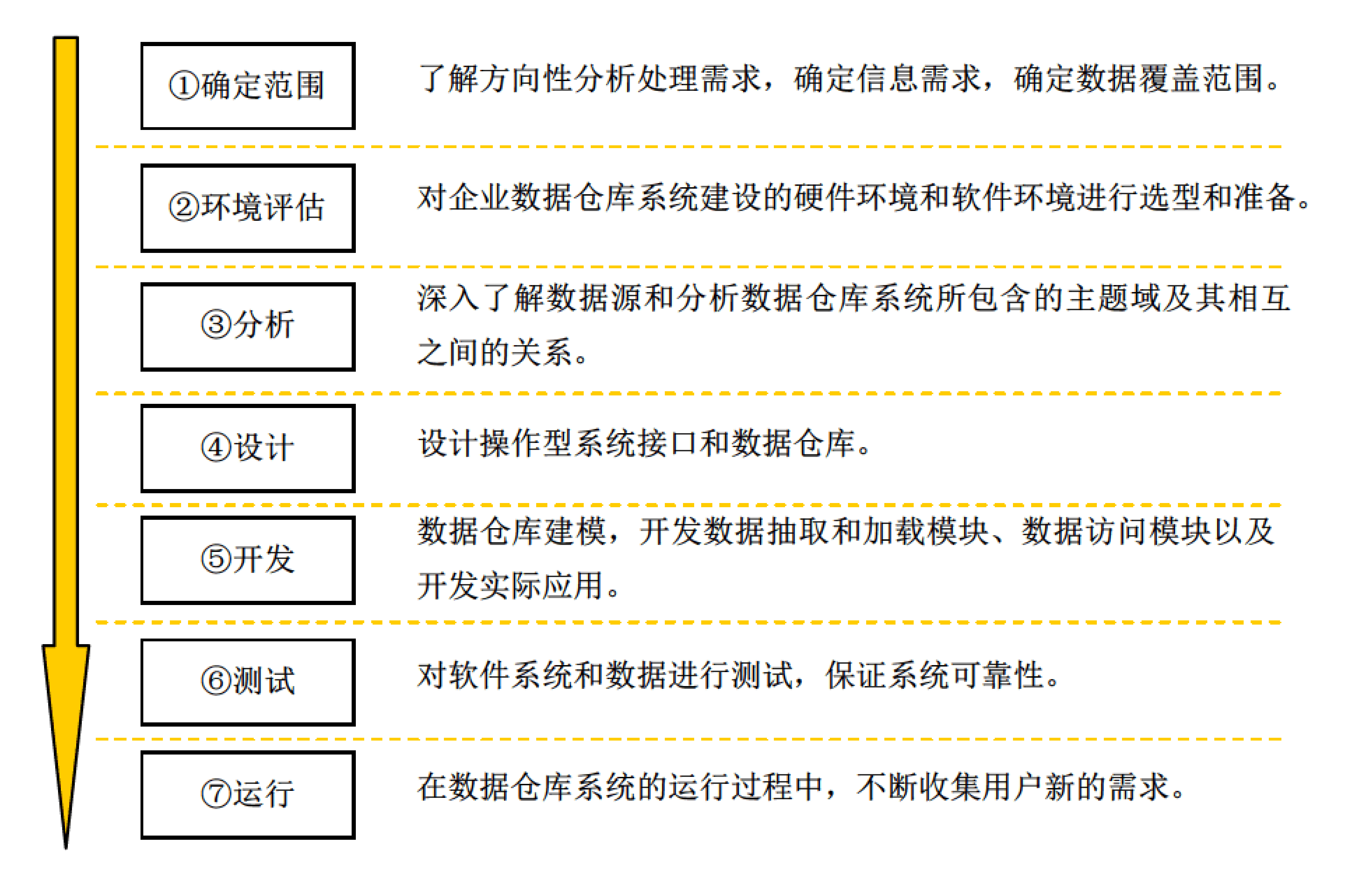


图 2-14　财务智能仓系统（BI-FIW）的工作流程

【案例解析】：在本案例中，创建帝高羊绒的数据仓库是一个庞大的系统工程，需要企业不断地去建立、发展和完善。

因此，企业可以首先提出一个全面、清晰的远景规划及技术实施蓝图，将整个项目的实施分成若干个阶段，并以“立体建模、分部解析、过程评估”为原则。做到这些，企业不仅可迅速地从当前投资中获得收益，而且可以在已有的基础上，结合其他已有的业务系统，逐步构建起完整、强大的数据仓库系统。

2.4.5　【案例】沃森人工智能计算机

日前，IBM 公司研发的电脑“沃森”战胜了美国电视智力节目《危险边缘》的两名人类选手，一时间，很多人担心，电脑越来越像人了，将会超越人类智慧。

沃森智能计算机是一台以 IBM 创始人托马斯 • 沃森名字命名的电脑，如图 2-15 所示。在硬件方面，IBM Power 7 系列处理器是当前 RISC 架构中最强的处理器——采用 45nm 工艺打造的 Power 7 处理器拥有 8 个核心 32 个线程，主频最高可达 4.1GHz，二级缓存更是达到了 32MB。在软件方面，IBM 研发团队为“沃森”开发的 100 多套算法可以在 3 秒内解析问题，检索数百万条信息然后再筛选还原成“答案”并以人类语言输出。

近日，IBM 又宣布将把“沃森”应用于云环境的开发平台，开放 API（Application Programming Interface，应用程序编程接口），让企业能够开发自家的“沃森”App，从而构建起“沃森”生态圈，将“沃森”应用到更广泛的领域。



图 2-15　沃森智能计算机

此外，IBM 还建立了一个“沃森”内容库，供应商可以为沃森提供内容，包括通用和专用的信息，如医疗保健等。“沃森”的优势是给出准确与可靠的答案，因此可以为医生提供更适合病人的解决方案。在医疗领域的应用将是“沃森”商用最主要的领域。

专家提醒

笔者认为，“沃森”项目如果想在医疗行业推行的话，还需要面临法律层面的问题。如果“沃森”诊断出错，而医生又听从了错误的诊断，那么“沃森”就会面临被患者告上法庭的危险，这对 IBM 而言是一个正在考虑的应用问题。

【案例解析】：目前，各行各业的数据资料都是以自然语言编写的，例如医疗行业的医疗记录、文本、杂志和研究资料，这些都是计算机难以理解的语言。另外，在零售、旅游、金融、电信、服务等行业，同样存在着大量以自然语言存储和编写的资料，如果存在一套能够在这些自然语言资料中快速找出准确答案的系统，将为行业带来巨大的改变。然而，本案例中的“沃森”具有理解自然语言、找到证据、判断这三大能力，这种“认知计算”能力让“沃森”在当前的大数据浪潮中大有用武之地。

“沃森”的工作过程实际上是一个完整的大数据分析过程：识别理解自然语言是处理非结构化数据的过程，找到证据就是从不同来源的大数据中检索的过程，判断就是给证据评分，作出最佳决策的过程。因此可以预见，“沃森”在大数据领域会有非常光明的前景。目前看来，沃森至少能在以下行业领域有所应用：电子、能源与电力、政府、卫生保健、保险、石油天然气、零售、通信、交通、银行与金融市场等。

3

架构：大数据基础设施

学前提示

大数据都会有自己的基础架构平台，一般推荐是基于云计算的动态弹性平台，因为它将为大数据的分析处理提供强有力的支撑。但是，企业要想让如此规模的数据真正转化为财富，数据中心必然将面临一次漫长而充满艰辛的基础设施及架构变革。

要点展示

- 探索全球，10 大大数据部署方案
- 掘金红海，10 大大数据分析平台
- 大数据基础设施应用案例

3.1 探索全球，10 大大数据部署方案

就在近两年，大数据应用突然爆炸，五彩缤纷的创意都变成现实。即使最谨慎的观察家也承认，大数据的商业应用时代已经来临，这都源于它前所未有的“从海量到精准”的预测能力。因此，大数据被认为是下一个创新、竞争和生产力的前沿，谁率先抓住大数据的先机即意味着能够在未来市场竞争中取得标杆地位。

俗话说：“工欲善其事，必先利其器。”在大数据实践之中，基础架构就犹如基石一般，是构建一切的基础，基础架构基石不稳，大数据“大厦将倾”，具有优秀的基础架构才能够让用户在未来的大数据之路中越走越宽。本节笔者就带大家一同回顾在世界各地那些不为人知却实际存在的大数据基础设施部署方案。

3.1.1 Netflix：掌握视频大数据炼金术

Netflix 是一家在线影片租赁提供商，能够提供超大数量的 DVD，而且让顾客可以快速方便地挑选影片，同时免费递送。

Netflix 已经成为美国国内规模最大的商业视频流供应商——目前拥有 2900 万视频流客户。这家公司同时也成为吸收新增数据的“海绵”——用户在看什么、喜欢在什么时段观看、在哪里观看以及使用哪些设备观看，爆增的信息量成为 Netflix 手中的宝贵资产。他们甚至掌握着用户在哪个视频的哪个时间点后退、快进或者暂停，乃至看到哪里直接将视频关掉等信息。

IHS 研究公司表示，2011 年 Netflix 的网上电影营收超过苹果，网络电影销量占据美国用户在线电影总销量的 45%，这主要得益于网络用户对在线视频的强大需求。

在美国众多的视频服务商里，Netflix 是最早尝试将大数据和媒体行业结合起来的公司。现在 Netflix 公司开始推出自己的原创节目，而节目制作的依据正是刚刚提到的这些数据。例如，Netflix 最新投资的电视剧“House of Cards”（纸牌屋），让人们见识了大数据分析对 Netflix 这样的新媒体公司的价值。

现在的 Netflix 不只提供线上影片出租与影片推荐服务，更是一家能够推出自制影集的全方位娱乐公司，其商业模式主要有两点，如表 3-1 所示。

表 3-1　Netflix 的商业模式

商业模式	主要特点
DVD 邮寄出租服务	打破原先的单片出租模式，改成创新的月租式服务，没有到期日也没有延迟罚款，消费者再也不用担心还片的问题。当消费者在线上选好想看的影片后，Netflix 便会运用其配送网络，在一天内寄出

续表

商业模式	主要特点
线上影片推荐系统	利用数据分析，根据消费者过去的影片评价，预测消费者接下来会想看什么样的影片，因此 Netflix 发展出 Cinematch 影片推荐引擎（Video Recommendation Engine），其运用 Big Data（大数据）和 Data Mining（数据挖掘），为消费者推荐影片

当初，Netflix 由于缺乏相应的设计人员和数据平台，因此颁发了 100 万美金大奖，希望世界上的计算机专家和机器学习专家们能够改进 Netflix 推荐引擎的效率。随后，来自 186 个国家的四万多个团队经过近 3 年的较量，一个由工程师、统计学家、研究专家组成的团队夺得了 Netflix 的大奖，该团队成功地将 Netflix 的影片推荐引擎的推荐效率提高了 10%。Netflix 大奖的参赛者们不断改进影片推荐效率，Netflix 的客户已经为此获益。

根据 Sandvine 市调公司研究报告，其下载量占全美网络下载量的 32.25%，以绝对优势占据第一名的位置，如图 3-1 所示。

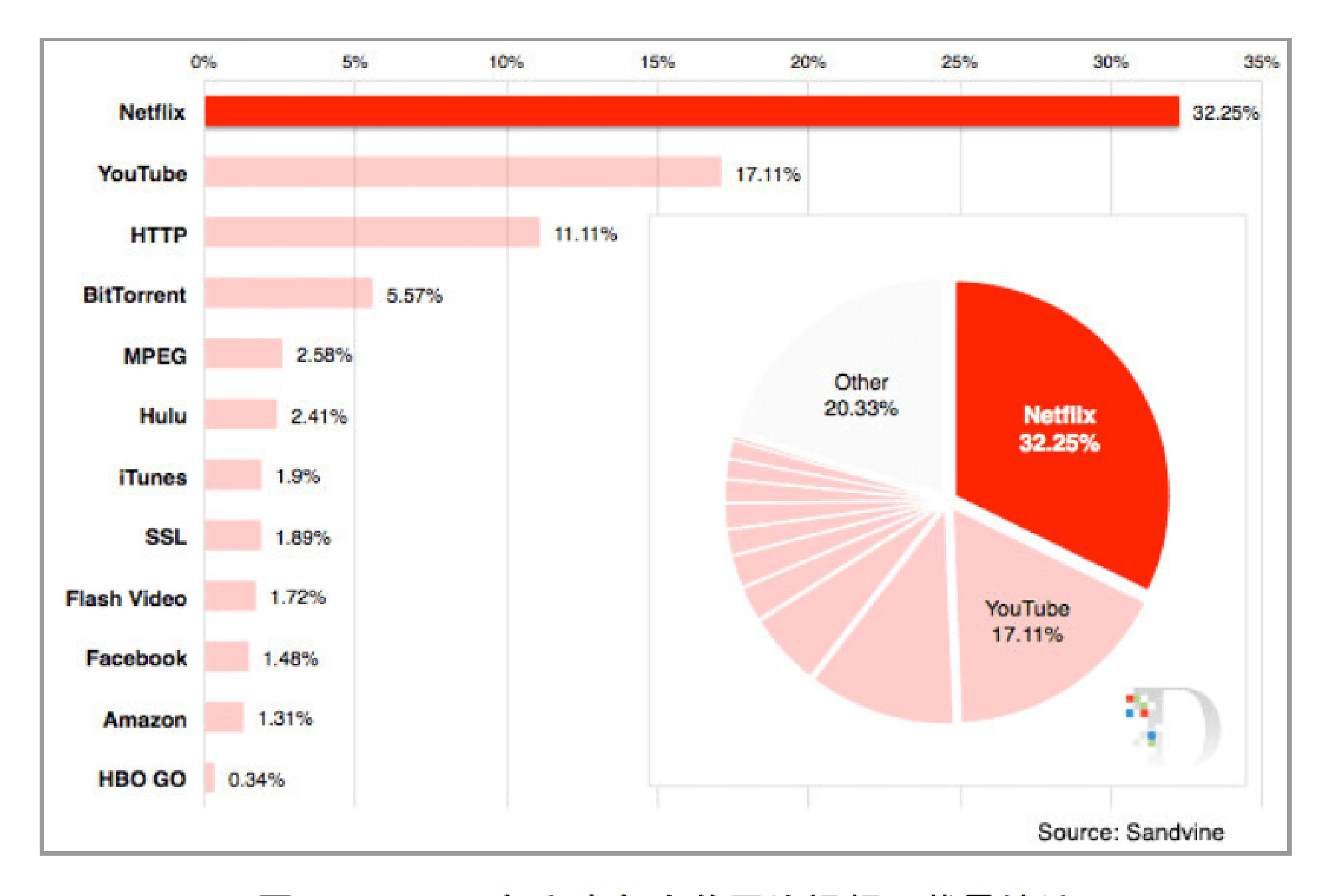


图 3-1　2013 年上半年全美网络视频下载量统计

专 家 提 醒

Netflix 在全球拥有超过 2500 万用户，每日平均 3000 万次的点击、播放、暂停、快转、回播，400 万次的评价行为，300 万次的搜索动作。

3.1.2　家谱网：建立更准确的血缘关系

家谱网到底有何魅力，先看看下面的两个资料。

资料 1：著名主持人马丁是马英九的远房亲戚，且比马英九长 6 代——两人均出自扶风马氏，赵国大将军赵奢（马服君）之后。马丁是赵奢的第 65 世孙，而马英九是赵奢的第 71 世孙。

资料 2：一个是中国奥运历史上首位冠军的安徽人许海峰，一个是来自台北的音乐人许常德，两位相隔几千公里的许姓男人，却有着一位共同的显赫祖先——唐朝宰相唐敬宗。

这些信息来自于 2008 年在国内上线的家谱网（jiapu.com），它是美国家谱网站 Ancestry 的中国版。Ancestry.com（家谱网）是一家家谱在线服务网站，拥有 10PB 的家族遗传数据，如图 3-2 所示。



图 3-2　Ancestry.com（家谱网）主页

长久以来，Ancestry.com 都是使用 apache Hadoop 以及其他的开源工具来进行数据 处理和分析的。然而，想要将 Hadoop 架构与 dba 数据处理联系起来，就极具挑战性，其中之一就是团队建设。因此，Ancestry.com 构建了自己的搜索引擎，并对算法以及记录连接软件进行了仔细的调优，该引擎可以对网站的结构化数据和非结构化数据进行遍历。

Ancestry.com 网站包含了大量出生、死亡、人口普查以及其他相关记录，这些记录起初大多是非结构化数据。随着用户以及家族数据的不断增长，Ancestry.com 公司希望改善其信息检索的算法。

不久后，公司招募了一些数据科学家，他们选择使用最新的工具，把 Hadoop、mapreduce 以及 R 语言引入了 Ancestry.com 的工具集。Ancestry.com 的团队使用 Hadoop 架构来对搜索进行优化，同时对客户流失率进行预测建模，并开始使用 Hadoop 以及相关的 hbase nosql 列式数据存储来对 Ancestry DNA 产品进行扩展。新的大数据

平台利用高级内容处理技术对全部相关信息加以索引，使用染色体 DNA 测试技术来为用户提供更好的服务，从而保证数据的可搜索性，甚至能够对远亲进行准确识别，从而让 Ancestry. com 获得用户的认可。

例如，Ancestry.com 通过对唾液进行采样，能够对客户的 DNS 进行排序并将结果与数据库中的其他客户加以匹配，客户甚至可以找到多年没有联系的表亲。

专家提醒

目前，家谱网累积的华人家谱总库中，包含 65584 种家谱数据，年代跨越明、清、民国以及当代，地域覆盖 24 个省及地区。其中，最早能追溯到 1498 年（明代）休宁陪郭（地名）的叶氏世谱。

Ancestry.com 帮助人们将自己与家庭史结合起来并创建独一无二的树状家谱。从表面上看，这个主意似乎没什么技术含量，但为了实现这项功能，网站需要维护超过 110 亿条记录与高达 4PB 的数据量——其中包括历史记录、出生记录、死亡记录、战争与移动记录甚至年鉴等，其中不少往往采取手写格式。

想要构建这一大数据平台，需要涉及大量的操作，大约有 70 万个 DNA 样本要与 Ancestry.com 数据库汇总已有的相同数量样本进行配对比较。Ancestry.com 的团队对学术算法进行了改写，从而可以在 Hadoop 和 hbase 上运行并行的任务，这样做可以大大提升海量数据处理的速度。

Ancestry.com 拥有明晰的盈利方式以及庞大的付费用户。付费用户可以分为两类，查看美国本土资料的用户和查看世界资料的用户，但收费不同。另外，在开发个人用户价值之外，Ancestry.com 还盯上了企业用户，例如数据库能使得企业的宣传销售更具针对性，以便提供个性化服务。数据库里的庞大家谱相当于“商品”，用户有需要时，便可付费购买。

3.1.3 西奈山：更深刻地理解数据形态

西奈山医院始建于 1852 年，是美国历史最悠久和最大的教学医院之一，以其在临床治疗、教学和科研方面的杰出成绩而闻名于世。

西奈山医院的很多新设备都是用来采集分析数据的，它运行 Hadoop 软件进行大数据分析。医院希望计算机专家利用大数据来寻找联系，例如在 ICU 中发现的微生物的 DNA，或者跟踪那些使用家用监控器的病人发来的数据流。

来自 Facebook 的首席数据科学家杰夫 • 哈默巴赫尔（Jeff Hammerbacher）负责设计这一切，他用分析目标在线广告的数据技术来分析各类基因数据和生物学信息，目的是减少医疗费用，同时探索“个性化医疗”。

目前，西奈山医院正利用来自大数据新兴企业 Ayasdi 公司的技术对整个大肠杆菌基

因组序列进行分析，其中包括超过 100 万个 DNA 变异，旨在努力理解某些菌株如何在与抗生素的共处中获得抗药性。细菌的抗药性影响着全球各地数以百万计的病人。Ayasdi 的技术为数学研究、拓扑数据分析（简称 TDA）开辟了一片新天地，有助于人们更深刻地理解数据形态。

在研究的基础上建立相应的数据库，结合日益普及的个人基因监测服务，正成为个性化医疗的基础。个性化医疗会彻底改变我们对待健康和疾病的方式，无论从政府、技术、学术还是产业层面，个性化医疗都是大势所趋。

3.1.4　CAISO：实现电厂电网的智能化

美国加利福尼亚州独立系统运营商（California Independent System Operator，CAISO）管理着全加州地区超过八成电网中的供电走向，每年提供的电力达到 2.89 亿千万时，惠及 3500 万民众，供电线路的总长度超过 25000 英里。

CAISO 所有的大型电厂都已经用上了企业后台办公系统，其中包括地理信息系统（GIS）、停电管理以及配电管理系统（DMS）。为了实现电网的智能化，CAISO 利用带有分析工具的历史数据功能接收数据流，将其与历史模式进行比较和对比，以便找出数据中的异常情况，如图 3-3 所示。

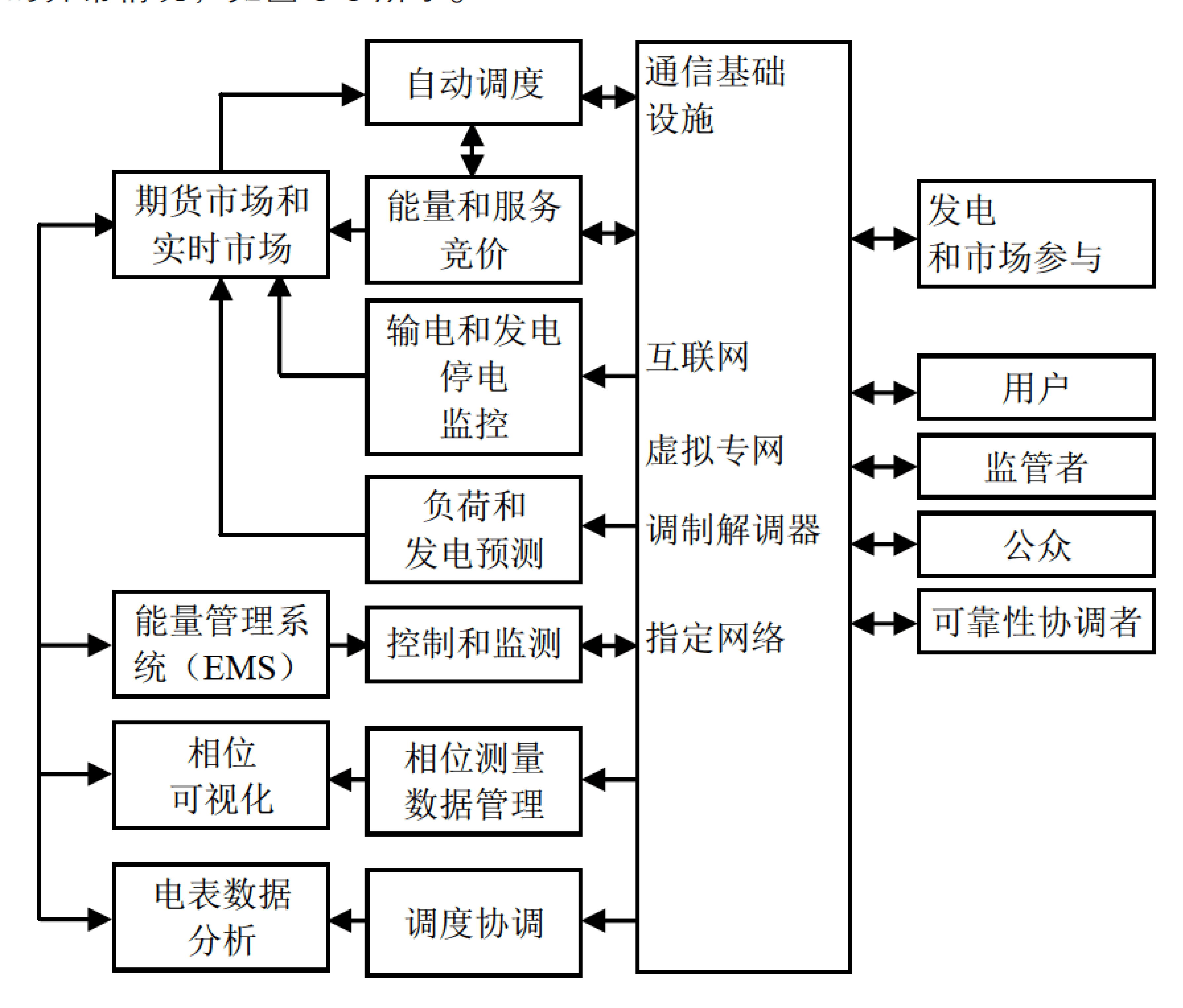


图 3-3　独立运营商（ISO）基础设施中的关键组件

ISO 利用 Space-Time Insight 公司的软件实现情景智能化机制，从而将来自多个来源的大规模数据进行关联与分析——其中包括天气状况、传感器数据以及计量设备测绘结果等，并以可视化形式帮助用户查看并理解如何对可再生能源进行优化，以实现整个电网的电力供需平衡并快速应对潜在危机。

3.1.5 Hydro One：把大数据放地图上

Hydro One（英语 Ontario，简称安省）是加拿大安大略省多伦多市最大的电力输送集团，负责为全省的家庭及企业提供电力。Hydro One 公司拥有并经营安大略省内总长达 29000 公里的高压输电网络以及总长达 123000 公里、直接面向 130 万用户的低压配电系统，如图 3-4 所示。

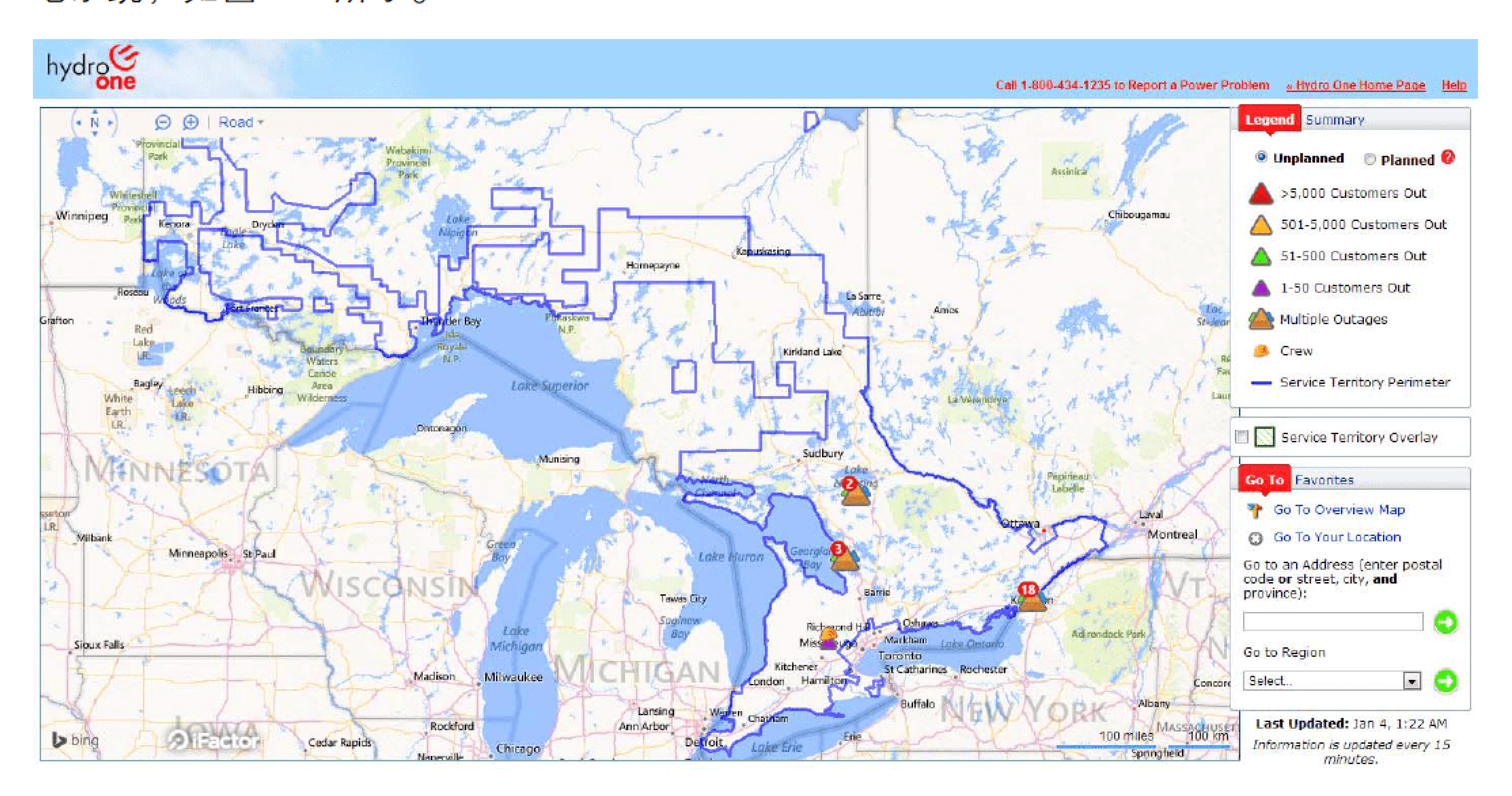


图 3-4　Hydro One 公司的高压输电网

Hydro One 使用的是由 Space-Time Insight 提供的地理空间与可视化分析软件，旨在改进当前输电与配电资产的健康性与可靠性。Space-Time Insight 是一家将大数据、数据可视化、地图 LBS 服务三者整合起来的公司，他们将企业需要的大量专业数据以地理信息的形式展现在地图上，让人们更好地了解、比较和研究他们所需的信息，如图 3-5 所示。

Space-Time Insight 打造的这套系统能帮助资产管理者及时获取相关情报，包括资产性能随时间推移而发生的变化、资产更换战略以及资产维护需求等。该方案还能将数据与其他多种不同系统的功能结合起来，包括 SAP ECC、SAP BW、GIS 系统以及实时数据等，从而帮助 Hydro One 对自身拥有的资产具备宏观掌控能力。

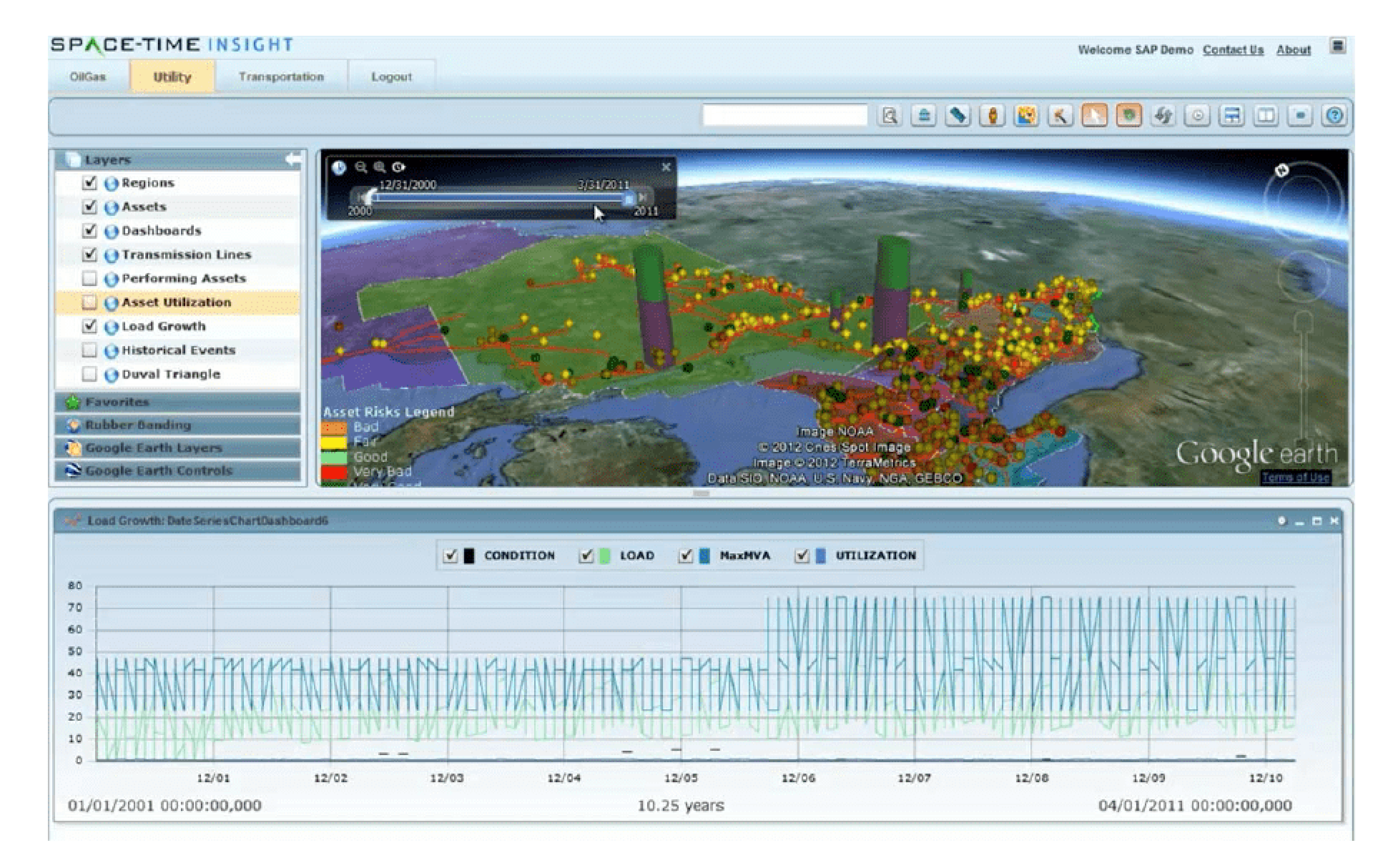


图 3-5　Space-Time Insight 的地理空间与可视化分析软件

虽然 Space-Time 的主要重心仍然放在电力行业，但无疑在其他能源、运输、气象等行业都有广阔的前景。而且除了企业市场，这类地图可视化技术在传统消费、生活服务市场也会有乐观的应用前景。

专 家 提 醒

在大数据时代，笔者认为企业更应该聚焦非结构化数据，结构化数据已经有了不错的归宿，非结构化数据才是我们处理的难题。据预测，到 2020 年，非结构化数据将数十倍于传统的结构化数据，成为大数据最主要的数据来源。

3.1.6　OHSU：结合数据虚拟化技术

俄勒冈健康与科学大学（Opegon Health and Science University，OHSU）是一所历史悠久、以研究为取向的最好的综合性公立大学，下辖两所医院、一座一级创伤恢复中心和一家儿童医院。学校致力于人类健康事业的发展，专注于提高食品安全、疑难疾病的预防与治疗等方面的研究。

为了追踪学校内 4000 个注液泵的实时位置与工作状态，更快地掌握注入到患者循环系统当中的液体、药物或者营养物质，校方将 Stanley Black 与 Decker Disivion Stanley Healthcare 提供的 Mobile View 软件与 Tableau 软件的数据虚拟化技术结合起来，改变传统的手动执行方式。该技术还允许校方对历史及当前资产数量进行分析，进而更好地规划未来数量水平，提高库存物资的分配与利用效率。

Tableau 公司将数据运算与美观的图表完美地结合在一起，如图 3-6 所示。它的程

序很容易上手，各公司可以用它将大量数据拖放到数字“画布”上，转眼间就能创建好各种图表。这一软件的理念是，界面上的数据越容易操控，公司对自己所在业务领域里的所作所为到底是正确还是错误，就能了解得越透彻。



图 3-6　Tableau Mobile 软件界面

专 家 提 醒

如今，每个企业都会有很多数据以及产生很多问题，为了分析这些数据，人们可以创建图表把数据与问题联系起来，但很多时候大家不确定从哪种图表可以得到自己要找的答案。Tableau 通过把数据搁置于独立的、静态的图中，限制了能够解决问题的范围。通过如何让数据成为决策的核心，以数据讲述一个故事来做出决策，以及添加一张图、提供过滤器以了解得更深入，Tableau 能帮助企业解决问题，它所带来的商业洞察力和回答问题的速度能与你的思想同步。

3.1.7　VTN：公共设施的实时 3D 模型

过去，大部分城市中的公共事业机构都是采用古老的手动记录方式，处理地下的各种资产，因此信息准确度十分低。例如，居民往往会由于某条供电线被意外切断或者某条供水管线老化爆裂而受到影响。

拉斯维加斯（Las Vegas）作为美国内华达州的最大城市，为了避免这些难题，市

政部门采取智能数据方式开发出一套实时公共事业网络模型。另外，VTN 咨询公司帮助市政当局通过各种渠道汇总数据，并利用 Autodesk 技术创建出实时 3D 模型。这套模型中包含着地上与地下的所有公共设施，目前已经被用于监测城市地下设施的具体位置以及运转状况。

专 家 提 醒

大数据虽然在不同的应用场景、不同的企业环境其应用方式会千差万别，但是常见的基本架构是大同小异的。经过分析与处理，能够应用于实践指导的信息数据会被整理到数据中心、应用程序以及基础设施当中，企业管理者需要以此为基础进一步将其导入各类系统及业务流程中，并最终获得（近乎）实时的决策能力。

3.1.8 戴德县：实现大型城市的智能化

迈阿密-戴德县（Miami-Dade County，Florida）是位于美国佛罗里达州东南部的一个县，2005 年估计人口达 2376014，成为美国的第 8 大县。

迈阿密-戴德县响应 IBM 提出的智能化城市倡议，希望将 35 个区域自治单位与迈阿密市聚拢起来，以便做出更为明智的管理决策——包括充分利用水资源、减少交通拥堵以及改善公共安全等，如图 3-7 所示。



图 3-7　智能化城市的构成体系

为此，IBM（国际商业机器公司，International Business Machines Corporation，IBM）通过云计算环境下的深层分析为该县带来一套情报仪表板，从而帮助各机关与部门彼此协作并实现可视化管理。

智慧城市具有 3 项基本特征，分别是物联化、互联化和智能化。基于这 3 个特征的 IBM 智慧地球计划自 2008 年开始展开，并且在近年来加速，且出现了很多成功的落地

项目。以 2012 年为例，IBM 先后发布了智慧云上的智慧交通新版本、智慧云上的智能运算中心新版本及智慧云上的智慧水利新版本。基于这一系列方案，IBM 搭建了涵盖公共安全、交通、水利等多个领域的解决方案，并搭建了智能运营中心。

专家提醒

笔者认为，城市管理只有利用大数据，才能获得突破性改善，诸多产业利用大数据，才能发现创新升级的机会点，进而获得先发优势……有了云计算、物联网，但缺乏大数据分析处理的核心技术，智慧城市的“大脑”就不够发达，“智商”就不够高，“能力”就不够强。

3.1.9 澳网：利用大数据分析做出决策

澳大利亚网球公开赛（Australian Open，简称“澳网”）是网球四大满贯赛事之一，也是四大满贯赛事中每年最先登场的，通常于每年 1 月的最后两个星期在澳大利亚墨尔本市的墨尔本公园举行。

澳大利亚网球公开赛的总奖金在 2013 年达到 3100 万澳元（3260 万美元），是四大满贯中奖金最高的赛事。澳大利亚网球公开赛自 1905 年创办以来，至今已经走过了一百多年的历史，赛事目前由澳大利亚网球协会（Tennis Australia）主办。

在平时，澳大利亚网球协会的运作状态与普通的小型企业没什么差别，然而一旦到了为期两周的澳网公开赛时期，协会瞬间就成了一家规模庞大、对数据极度渴求的大型企业——他们需要不间断地访问准确内容、数据以及统计结果，从而进行分析并做出决策。

下面提供一组 2013 年度澳大利亚网球公开赛的统计资料：

- 684457 名球迷到现场观看了比赛。
- 澳网网站有 1410 万绝对造访人次。
- 澳网 Social Leaderboard 追踪到 900 多万涉及球员的 Twitter。
- 澳大利亚网球协会在比赛期间获取了约 60TB 的数据和视频资源，本次赛事男子抽签 127 场比赛打了 764 盘。

目前，澳大利亚网球协会采用 IBM 的实时数据分析软件来检查赛程进行状态、运动员人气、历史数据记录以及社交媒体上球迷们对比赛网站提出的数据需求。根据实际需求，这项技术能够为分析工作分配必要的计算资源。

澳大利亚网球公开赛网站上提供 IBM SlamTracker 工具，用以分析 8 年大满贯赛事比赛的 4100 万个数据点，如图 3-8 所示。除了其他方面之外，该工具还有一项功能，称为“Keys to the Match”，可帮助球迷了解球员为了在某项特定比赛中取胜，需要做哪些工作。当一场比赛拉开帷幕时，该工具根据关键点测评每个球员的表现并实时更新，从而提供更深入的洞察力，包括高比例第二发球接发或者上网成功率是否有助于挑高球

过人。

例如，在李娜与小威廉姆斯的比赛中，李娜一方获得赢球的关键包括 3 个指标（如图 3-8 所示）：1. 一发（首次发球）得分率超过 69%；2. 4～9 拍相持中得分率要超过 48%；3. 发球局 30-30 或 40-40 时得分率要超过 67%。

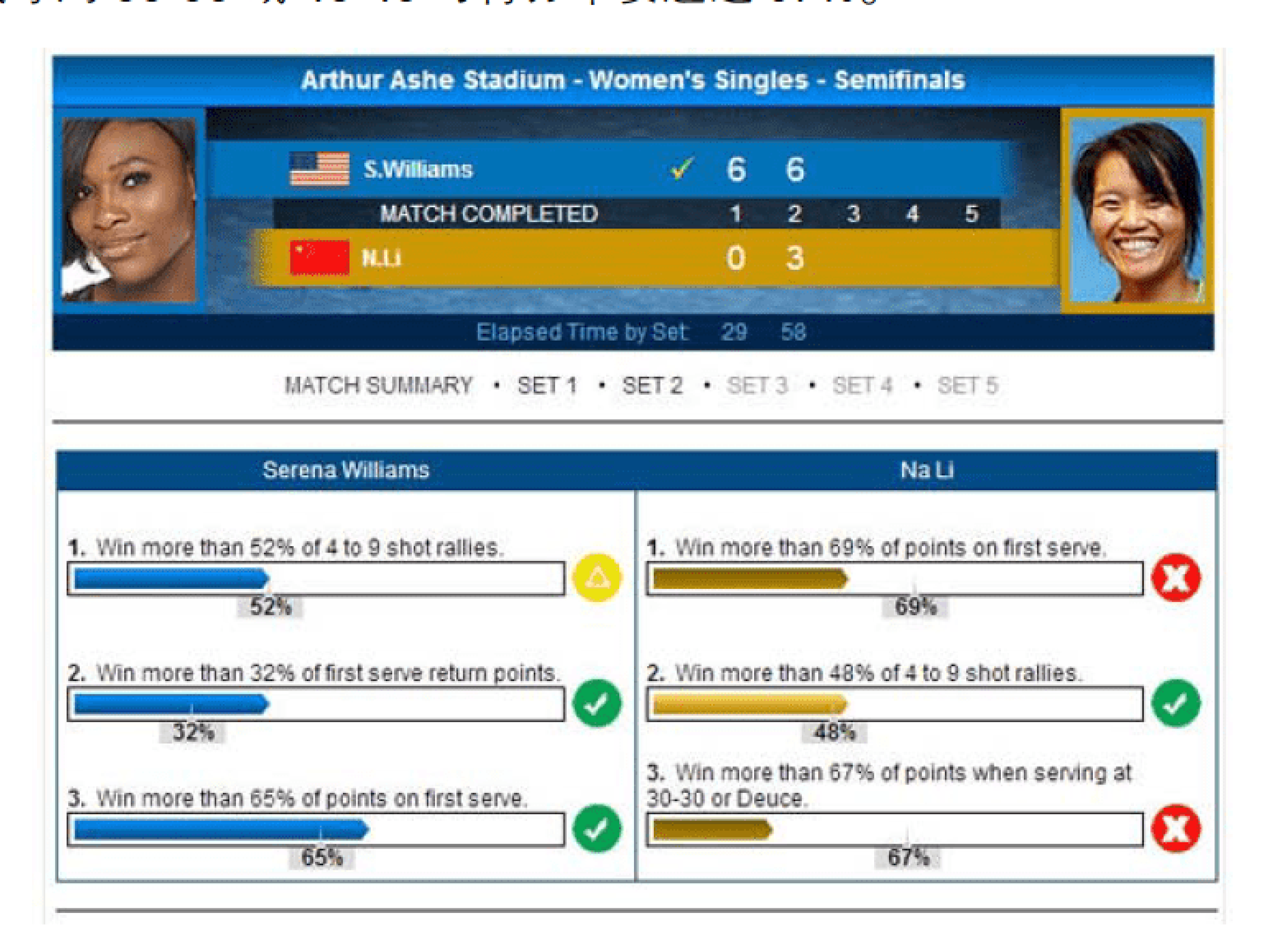


图 3-8　IBM 智能分析平台 SlamTracker

而在实际的比赛中，李娜只完成了第二项指标，相比之下，小威廉姆斯则完成了两个指标。因此，据此分析，李娜出局主要跟一发得分率低、双方平分时未能获得关键分数有关。

为了打造完善的大数据基础设施，澳大利亚网球协会还与 Aruba 共同构筑安全可靠、灵活、可扩展的无线网络，而它所具备的环境意识功能，更可有效地管理紧凑赛程网络状况。这意味着协会能够非常准确预测网络连接需求高峰的时间和地点，从而调整网络满足所需。

据悉，在 2013 年澳网比赛的两周内，单是#ausope 标签就有一百多万条微博，澳网 Facebook 页面增加到约 887158。社交媒体洞察力在澳大利亚网球协会和其他机构的决策以及与客户互动方面，具有越来越重要的作用。在该满贯赛事期间，使用先进的 IBM 分析软件和自然语言处理技术来评估 Twitter、Facebook、新闻网站、博客和视频等网站上数十万社交媒体消息分享的正面和负面情绪。

专 家 提 醒

数据分析已经深入体育运动，并且在改变体育运动的发展模式。大数据将改变我们消费、观看网球等体育运动以及与其进行互动的方式。那些拥护并利用该技术为业务决策以及与球迷联络提供相关信息的机构，和竞争对手相比，将赢得竞争优势。

3.1.10 DPR：结合 3D 技术与大数据

美国加州大学旧金山分校斥资 15 亿美元在米慎湾兴建了一座医学中心，这也是第一座建造时间超过十年的医学中心，承包商为 DPR Construction 公司。

DPR Construction 公司利用 Autodesk 公司的 3D 技术，帮助设计师们收集空气流量、建筑物朝向、楼体间距、环境永续性以及建筑性能等数据，并将结果导入到一套单独的虚拟模型当中。通过这种方式，建筑师、设计师以及旗工队伍能够以可视化方式掌握遍布整个运作环境下的数亿个数据标记。

专 家 提 醒

Autodesk 公司的 Vault 数据管理软件可以帮助设计、工程和施工团队组织、管理和跟踪数据创建、仿真和文档编制流程。借助版本管理功能，企业可以更好地控制设计数据，快速查找和重用设计数据，从而更加轻松地管理设计与工程信息。使用 Autodesk Vault 后，用户可以在一个平台下管理所有的 CAD 和非 CAD 数据，从而提高工作效率，如图 3-9 所示。

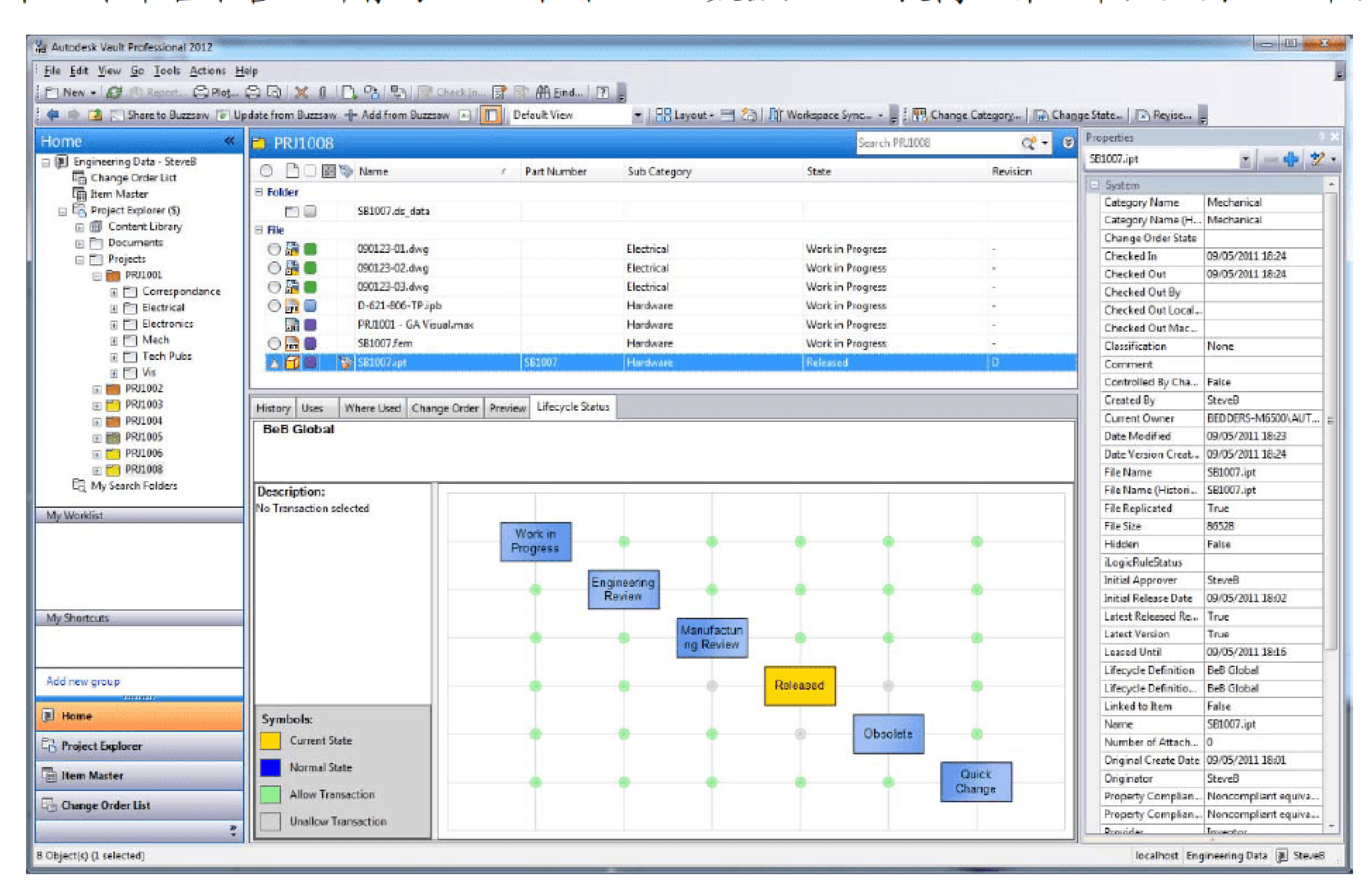


图 3-9 Autodesk Vault

3.2 掘金红海，10 大大数据分析平台

“大数据”近几年来可谓蓬勃发展，它不仅是企业趋势，也是一个改变了人类生活的技术创新。在大数据的帮助下，警察可以通过犯罪数据和社会信息来预测犯罪率，部

分科学家通过遗传数据预测疾病的早期迹象。可以说，现在整个行业都非常看好大数据。

毫无疑问，在大数据时代下，企业和机构要想实现更大的业务价值，首先需要解决的就是基础架构问题，基础架构之中存储又是重中之重。目前，我国从事大数据领域的企业少之又少，而国外的科技企业将大数据看作是云计算之后的另一个巨大商机，很多企业开始加入到大数据的淘金队伍中，这一领域已经成为实实在在的红海。

本节将介绍全球 10 大著名的大数据分析平台（注意：排名不分先后），他们是大数据领域的“时代先锋”，他们都看到了大数据带来的大机会。

3.2.1 IBM：大数据领域的传统巨头

企业名称：IBM（如图 3-10 所示）

分析平台：InfoSphere 大数据分析平台

上线时间：2011 年 5 月

公司地址：美国纽约州阿蒙克市

企业网址：http://www.ibm.com/

主要业务：软件、服务器、存储、IT 服务以及云计算等解决方案

业务方向：主要面向大企业等

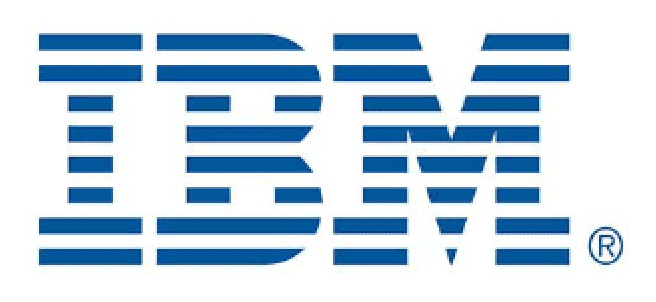


图 3-10　IBM Logo

IBM 是一个拥有 101 年历史的公司，总部在美国东海岸。它曾经生产打字机，还生产大型 PC 机，其产品使用开源技术进行交互操作。在 IBM 的发展过程中，很多产品都是通过一系列兼并得来的。最重要的是，IBM 是一家服务公司，有着工作在全球各地的顾问团队。

IBM 向我们展示了将大数据与企业连接的重要性和一个主流服务组织，它还展示了向业务软件中嵌入分析功能的力量。

2011 年 5 月，IBM 正式推出 InfoSphere 大数据分析平台。InfoSphere 大数据分析平台包括 BigInsights 和 Streams，二者互补。

- BigInsights 基于 Hadoop，它对大规模的静态数据进行分析，提供多节点的分布式计算，可以随时增加节点，提升数据处理能力。例如，丹麦能源企业维斯塔斯（Vestas）通过使用 BigInsights 大数据软件分析 PB 字节级别的天气数据，改善风力涡轮机的放置位置，从而获得最佳能量输出效果——以前需要数周方可完成的分析现在仅需不到一个小时。

专 家 提 醒

Hadoop 本身不提供分析的功能，因此 BigInsights 平台增加了文本分析、统计分析工具。

- Streams 采用内存计算方式分析实时数据。Streams 最早是美国国土安全部和 IBM 合作的项目，国土安全部出于反恐目的，需要实时分析电话语音信息，这个项目最终发展成为一个商用的项目。

另外，InfoSphere 大数据分析平台还集成了数据仓库、数据库、数据集成、业务流程管理等组件。

3.2.2 亚马逊：完美结合大数据与云

企业名称：亚马逊（如图 3-11 所示）

分析平台：弹性 MapReduce（Amazon Elastic MapReduce）

上线时间：2009 年

公司地址：美国华盛顿州西雅图

企业网址：http://www.amazon.com/

主要业务：电子商务、云服务

业务方向：主要面向大企业等市场

图 3-11　亚马逊 Logo

亚马逊的老本行是图书音像制品销售，但现在这只是其业务的一个组成部分，而且已经不是公司业务的核心。如今，亚马逊已经成为一家拥有大数据，并以此获得持续利润的云计算企业。电子商务的数据，合并在这些大数据之中，仅仅是亚马逊将数据变为现金的一种方式。

亚马逊对于云计算和大数据具有先见之明，早在 2009 年就推出了“弹性 MapReduce（Amazon Elastic MapReduce）”系统。MapReduce 本身是一种编程模型，用于大规模数据集（大于 1TB）的并行运算，常用作 Web 索引、数据挖掘、日志文件分析、金融分析、科学模拟和生物信息研究等。

然而，“弹性 MapReduce”是一项能够迅速扩展的 Web 服务，其运行在亚马逊弹性计算云（Amazon EC2）和亚马逊简单存储服务（Amazon S3）上。面对数据密集型任务，例如互联网索引、数据挖掘、日志文件分析、机器学习、金融分析、科学模拟和生物信息学研究，用户需要多大容量，“弹性 MapReduce”系统立即就能配置到多大容量。

对于 MapReduce，笔者认为可以将其简单理解为：把一堆杂乱无章的数据按照某

种特征归纳起来，然后处理并得到最后的结果。

专家提醒

亚马逊的“弹性 MapReduce”服务系统是在 AWS 平台（AWS Enterprise BPM Platform，业务流程管理开发平台）之上的 Hadoop 实现，它用来简化新的 MapReduce 应用，从而让这项技术拥有更加广大的受众。

3.2.3 甲骨文：高集成度大数据平台

企业名称：甲骨文（如图 3-12 所示）
分析平台：Oracle 大数据机
上线时间：2010 年
公司地址：美国加利福尼亚州红木滩
企业网址：http://www.oracle.com/
主要业务：数据库、应用软件以及相关的咨询、培训和支持服务
业务方向：主要面向大企业等市场

ORACLE®

图 3-12　甲骨文 Logo

甲骨文公司，全称甲骨文股份有限公司，是全球最大的企业软件公司，也是继 Microsoft 及 IBM 后全球收入第三多的软件公司。

伴随大数据而至，大数据分析和管理得当与否将对企业数据中心产生极大影响。作为全球最大数据库软件公司，甲骨文应时而行，推出针对大数据的众多技术产品来满足企业需求，同时提升自身的价值。

2011 年 10 月，甲骨文正式推出了 Oracle 大数据机（Oracle Big Data Appliance）为许多企业提供了一种处理海量非结构化数据的方法。尤其是对于那些正在寻求以更高效的方法来采集、组织和分析海量非结构化数据的企业而言，该产品具有很大的吸引力。

Oracle 大数据机同 Oracle Exadata 数据库云服务器、Oracle Exalytics 商务智能云服务器和 Oracle Exalogic 中间件云服务器一起组成了 Oracle 最广泛的高度集成化系统产品组合，其可以帮助客户获取和管理各种类型的数据，并且可结合现有企业数据来分析，获得新的见解，从而帮助客户在充分获取信息的情况下做出最恰当的决策。

专家提醒

Oracle 大数据机能够拥有强大优化企业数据仓库的能力，主要源自其配备有 Oracle Big Connectors 软件。Oracle 大数据机旨在帮助客户利用 Oracle 数据库 11g 便捷整合存储在 Hadoop 和 Oracle NoSQL 数据库中心的数据。

3.2.4 谷歌：价值无可估量的大数据

企业名称：谷歌（如图 3-13 所示）

分析平台：BigQuery

上线时间：2011 年

公司地址：美国加利福尼亚州山景城

企业网址：http://www.google.com/

主要业务：互联网搜索、云计算、广告技术

业务方向：面向各类企业市场



图 3-13 谷歌 Logo

Google 在搜索界的地位是无人能及的。但是，Google 的产品和服务早已不仅仅局限于搜索。如今，Google 的产品包括广告（AdWords）、交流和分享（Drive 和 Hangouts）、开发资源（OpenSocial）、社交网络（Google +）、地图（Google Maps）、流媒体（Google Play）、统计工具（Analytics）、操作系统（Android 和 Chrome OS）、桌面和移动应用（Gmail）以及硬件（Galaxy Nexus）。因此，如果对其拥有的海量数据进行深入挖掘，这对于提升谷歌搜索乃至所有谷歌服务的价值无可估量。

BigQuery 是 Google 于 2011 年底正式推出的一项 Web 服务，通过该服务，开发者可以使用 Google 的架构来运行 SQL 语句对超大型的数据库进行操作。即 BigQuery 可以对开发者上传的超大型数据进行直接交互式分析，开发者无需投资建立自己的数据中心。据悉，BigQuery 引擎可以快速扫描高达 70TB 未经压缩处理的数据，并且可马上得到分析结果。

3.2.5 微软："端到端"大数据平台

企业名称：微软（如图 3-14 所示）

分析平台：PDW、SQL Server 2012 数据库平台

上线时间：2011 年

公司地址：美国华盛顿州雷德蒙市

企业网址：http://www.microsoft.com/

主要业务：电脑软件服务

业务方向：面向各类企业市场



图 3-14 微软 Logo

EMC、IBM 和甲骨文在 2011 年都大力追捧 Hadoop，于是微软也进入这个市场就不足为奇了。如今，微软已经具备了打造"端到端"的大数据平台的能力。

专 家 提 醒

“端到端”流程是从客户需求端出发，到满足客户需求端去提供端到端服务，端到端的输入端是市场，输出端也是市场。

2011 年初，微软发布了 SQL Server R2 Parallel Data Warehouse（PDW，并行数据仓库），PDW 使用了大规模并行处理技术来支持高扩展性，它可以帮助客户扩展部署数百 TB 级别数据的分析解决方案。

微软在 2012 年上半年正式发布了 SQL Server 2012 数据库平台，并添加了 Hadoop 的相关服务，逐渐将数据业务延伸到非结构化数据领域。而伴随 Windows Azure Marketplace 和 SharePoint 等工具的推出，微软已经具备了打造端到端的大数据平台的能力。

专 家 提 醒

Windows Azure Marketplace 将实现大数据的共享，通过开放数据协议（OData）展现数百种来自微软和第三方的应用程序和数据挖掘算法。用户还可以使用 SQL Server 分析服务（SSAS）的 Power Pivot 和 Power View，从结构化和非结构化数据中获得可执行的洞察力，通过微软提供的连接器就可以对 Hadoop 分布式文件系统中的非结构化数据进行分析与展现。

3.2.6 EMC：针对海量数据分析应用

企业名称：EMC（如图 3-15 所示）

分析平台：EMC Greenplum Unified Analytics Platform 大数据分析平台

上线时间：2011 年

公司地址：美国马萨诸塞州（麻省）Hopkinton 市

企业网址：http://www.emc.com/

主要业务：信息存储及管理产品、服务和解决方案

业务方向：面向各类企业市场

EMC²
where information lives®

图 3-15　EMC Logo

EMC 公司是全球信息存储及管理产品、服务和解决方案方面的领先公司。EMC 是每一种主要计算平台的信息存储标准，而且世界上最重要信息中的 2/3 以上都是通过 EMC 的解决方案管理的。

EMC 推出了全新 EMC Greenplum Unified Analytics Platform（UAP）平台，数据团队和分析团队可以在该平台上无缝地共享信息、协作分析。Greenplum UAP 是唯一的统一数据分析平台，可扩展至其他工具，其独特之处在于，它将对大数据的认知和分享贯穿于整个分析过程，实现比以往更高的商业价值。

随着 EMC Greenplum 统一分析平台的问世，EMC 提供关键技术帮助机构用户提取

大量数据的核心价值，并创造更多、更灵活、基于数据的业务机会。

专 家 提 醒

EMC 为大数据开发的硬件是模块化的 EMC 数据计算设备（DCA），它能够在一个设备里面运行并扩展 Greenplum 关系数据库和 Greenplum HD 节点。DCA 提供了一个共享的指挥中心（Command Center）界面，让管理员可以监控、管理和配置 Greenplum 数据库和 Hadoop 系统性能及容量。

3.2.7 英特尔：用 Hadoop 靠拢大数据

企业名称：英特尔（如图 3-16 所示）

分析平台：Hadoop 商业发行版（Apache Hadoop Distribution）

上线时间：2012 年

公司地址：美国加利福尼亚州圣克拉拉市

企业网址：http://www.intel.cn/

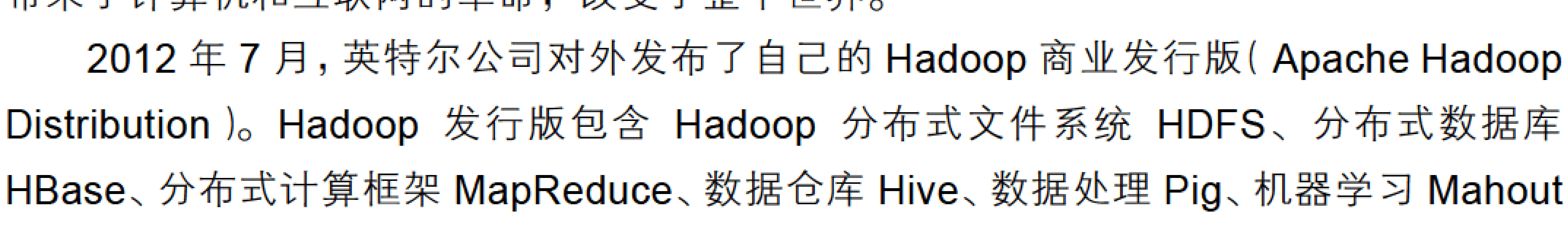
主要业务：客户机、服务器、网络通信、互联网解决方案和互联网服务

业务方向：面向各类企业市场



图 3-16 英特尔 Logo

英特尔公司是全球最大的半导体芯片制造商，成立于 1968 年。1971 年，英特尔推出了全球第一个微处理器，带来了计算机和互联网的革命，改变了整个世界。


2012 年 7 月，英特尔公司对外发布了自己的 Hadoop 商业发行版（Apache Hadoop Distribution）。Hadoop 发行版包含 Hadoop 分布式文件系统 HDFS、分布式数据库 HBase、分布式计算框架 MapReduce、数据仓库 Hive、数据处理 Pig、机器学习 Mahout 商业套件。

英特尔 Hadoop 发行版包含了所有的分析、集成以及开发组件，并对不同组合之间进行了更加深入的优化。此外，还添加了英特尔 Hadoop 管理器（Hadoop Manager），其从安装、部署到配置与监控，可以提供对平台的全方位管理。目前，英特尔已经开放了免费下载，随着推广力度的不断加大，相信英特尔的 Hadoop 还是能够很轻松地在国内大数据市场分一杯羹的。

3.2.8 NetApp：让大数据变得更简单

企业名称：NetApp（如图 3-17 所示）

分析平台：NetApp StorageGRID

上线时间：2011 年

公司地址：美国加利福尼亚州森尼韦尔

企业网址：http://www.netapp.com

主要业务：储存和数据管理解决方案

业务方向：面向各类企业市场



图 3-17　NetApp Logo

Network Appliance，Inc.（简称 NetApp，美国网域存储技术有限公司）是 IT 存储业界的佼佼者，自 1992 年创建以来，不断以创新的理念和领先的技术引领存储行业的发展。NetApp 公司倡导向数据密集型的企业提供统一的存储解决方案，用以整合网络上来自服务器的数据，并有效管理呈爆炸性增长的数据。

StorageGRID 是 NetApp 的对象存储平台，是一个久经验证的对象存储软件解决方案，设计用于管理 PB 级、全球分布的存储库，这些存储库包含企业和服务提供商的图像、视频和记录。通过消除数据块和文件中数据容器的典型约束，NetApp StorageGRID 提供了强大的可扩展性，它支持单个全局命名空间内的数十亿个文件或对象和 PB 级容量。NetApp 目前将 StorageGRID 产品并入其 E 系列，属于分布式内容存储类别。

NetApp 自创建以来，市场业务表现亦出众超群，公司一直保持了极高的成长率，并不断扩展用户群，其客户领域包括通信、金融、能源、政府、制造、教育及各类媒体、各种企业和服务提供商。

3.2.9　惠普：构建灵活的"智能环境"

企业名称：惠普（如图 3-18 所示）

分析平台：Vertica Analytics Platform、Information Optimization solutions

上线时间：2011 年

公司地址：美国加利福尼亚州帕罗奥多市

企业网址：www.hp.com

主要业务：打印机、数码影像、软件、计算机与资讯服务

业务方向：面向各类企业市场



图 3-18　惠普 Logo

惠普（HP）是一家业务机构遍及全球 170 多个国家和地区的科技公司。作为世界最大的科技企业，惠普提供打印机、

个人计算机、软件、服务和 IT 基础设施等产品，帮助客户解决问题。

2011 年，惠普子公司 Vertica 发布 Vertica Analytics Platform 大数据平台，意在帮助企业迅速洞悉关键的业务信息，辅助决策过程。Vertica Analytics Platform 能够让用户大规模实时分析物理、虚拟和云环境中的结构化、半结构化和非结构化数据，从而深入洞悉“大数据”。

2012 年 6 月，惠普发布信息优化解决方案（Information Optimization solutions），旨在帮助企业充分利用爆炸性增长的运营数据、应用数据和设备数据。

2013 年初，惠普推出了最新版本惠普 Vertica 分析平台 6.1（HP Vertica Analytics Platform 6.1），其能够对大数据进行简化。据了解，该平台将帮助企业通过分析包、性能提升、加强与 Hadoop 的集成以及简化 Amazon EC2 云部署，从而优化大数据并将其转化为利润。

另外，惠普还扩展了其业界领先的数字营销平台，发布了全新的 Autonomy 解决方案——Optimost Clickstream Analytics，其在电子商务中为市场营销人员提供客户访问、对话和参与情况的单一、连续的视图，为实现“瞬捷”企业构建灵活的智能环境。

专家提醒

在当今瞬息万变的商业环境下，“瞬捷”企业的创新优势在于能够提供与时俱进的、有竞争力的产品和服务，以加快业务增长，其优化特性则是指具备更高的投资回报率和更低的成本。

3.2.10 Sybase：彻底改变大数据分析

企业名称：Sybase（如图 3-19 所示）

分析平台：Sybase IQ

上线时间：2009 年

公司地址：美国加利福尼亚州 Dublin 市

企业网址：www.sybase.com

主要业务：应用平台、数据库和应用软件

业务方向：面向各类企业市场



图 3-19 Sybase Logo

Sybase 公司成立于 1984 年 11 月，是全球最大的独立软件厂商之一，致力于帮助企业等各种机构进行应用、内容及数据的管理和发布。Sybase 的产品和专业技术服务，为企业提供集成化的解决方案和全面的应用开发平台。

Sybase 公司推出的 Sybase IQ 是一款为数据仓库设计的关系型数据库。IQ 的架构与大多数关系型数据库不同，其特别的设计用以支持大量并发用户的即时查询。它的设计与执行进程优先考虑查询性能，其次是完成批量数据更新的速度。而传统关系型数据

库引擎的设计既考虑在线的事务进程又考虑数据仓库。

其中，Sybase IQ 15.4 是面向大数据的高级分析平台，它将大数据转变成可指挥每个人都行动的情报信息，从而在整个企业的用户和业务流程范围内轻松具备大数据的分析能力。

Sybase IQ 大大节约了数据存储成本，而且通过其强大的可扩展性为企业提供了灵活的选择。另外，IQ 比传统的数据库更容易维护，不需要经常的人工调优。简单的扩展实现以及快速的部署时间等，都大幅度地降低了企业开发数据仓库的成本。

3.3 大数据基础设施应用案例

目前，很多人只将眼光盯在数据分析与处理层面，而笔者认为，用户在尝试大数据解决方案之前，更应从全面角度去审视自身的基础架构是否适合大数据未来的需求与发展。简而言之，就是“大数据实践，基础架构先行”。只有如此，方能在大数据浪潮之中淘得金。本节主要介绍大数据基础设施的应用案例。

3.3.1 【案例】Streams 监控婴儿 ICU 感染

ICU 病室是医院主要科室之一，因其病人多来自于院内各科室，且病情危重，致使院内感染发生率在 ICU 相对增高。又因病人治愈后，又回散到原科室，使在 ICU 的耐药菌株被携带到医院各处而引起流行。由此可见，做好 ICU 病室的感染控制十分有必要。

安大略理工大学（UOIT）是加拿大最现代的公立大学，其拥有北美一流的教学设备和师资。学校目前正在使用 Streams 监控新生婴儿，提前 24 小时预测 ICU 感染。

安大略理工大学健康信息学首席科学家 Carolyn McGregor 博士称，这一技术让安大略理工大学能够搞清楚这些数据并分析它们，如揭示败血症的发生前兆，以及这些问题发生前的多种条件。

Streams 提供了一种操作系统实现这个功能，其在多台计算机之间共享一个特定程序，这样系统作为一个整体就可以在不把数据提交到硬盘的情况下生成答案，解决了针对能够实时处理生成的海量流数据的平台和架构的一种迫切需求。

【案例解析】：在本案例中，InfoSphere Streams 是一款满足即时处理、过滤和分析流数据需要的应用程序。流数据包括传感器数据（环保以及工业生产传感器产生的数据、监控视频、GPS 产生的数据等）、“数据废气”（如网络/系统/Web 服务器/应用程序服务器日志文件）、高速交易数据（如金融交易和呼叫详细记录）等。

预测分析与结构化数据未来将在医疗保健领域中被广泛应用，以帮助降低成本，防止病人病情恶化。大数据分析平台使医疗机构拥有更好使用这些信息的能力，这将从本质上改变医疗保健行业的未来。

3.3.2 【案例】沃尔玛打造商业数据中心

在 2012 年财政年度报表上，沃尔玛记录了 4440 亿美元的销售额，这个数字比奥地利的 GDP 多 200 亿美元。如果沃尔玛是一个国家的话，它将是第 26 个世界最大的经济体。

沃尔玛为何取得如此大的成就？笔者发现，沃尔玛其实是最早通过利用大数据而受益的企业之一，曾经拥有世界上最大的数据仓库系统。早在 2007 年，沃尔玛就已建立了一个超大的数据中心，其存储能力高达 4PB 以上。《经济学人》曾报道，沃尔玛的数据量已经是美国国会图书馆的 167 倍。

众所周知，沃尔玛的供应链是全球零售商中最先进的。早在 20 世纪 80 年代，沃尔玛就率先开发数据交换系统（Electronic Data Interchange，EDI）与供应商信息系统直接对接，实现了商品的自动补货。如图 3-20 所示为基于 EDI 的供应链信息组织与集成模式。为了加强数据的共享，沃尔玛还投资 4 亿美元发射卫星进行全球数据联网。通过全球网络，沃尔玛数千家门店可在一小时内对每种商品的库存、在架以及销售盘点一遍。

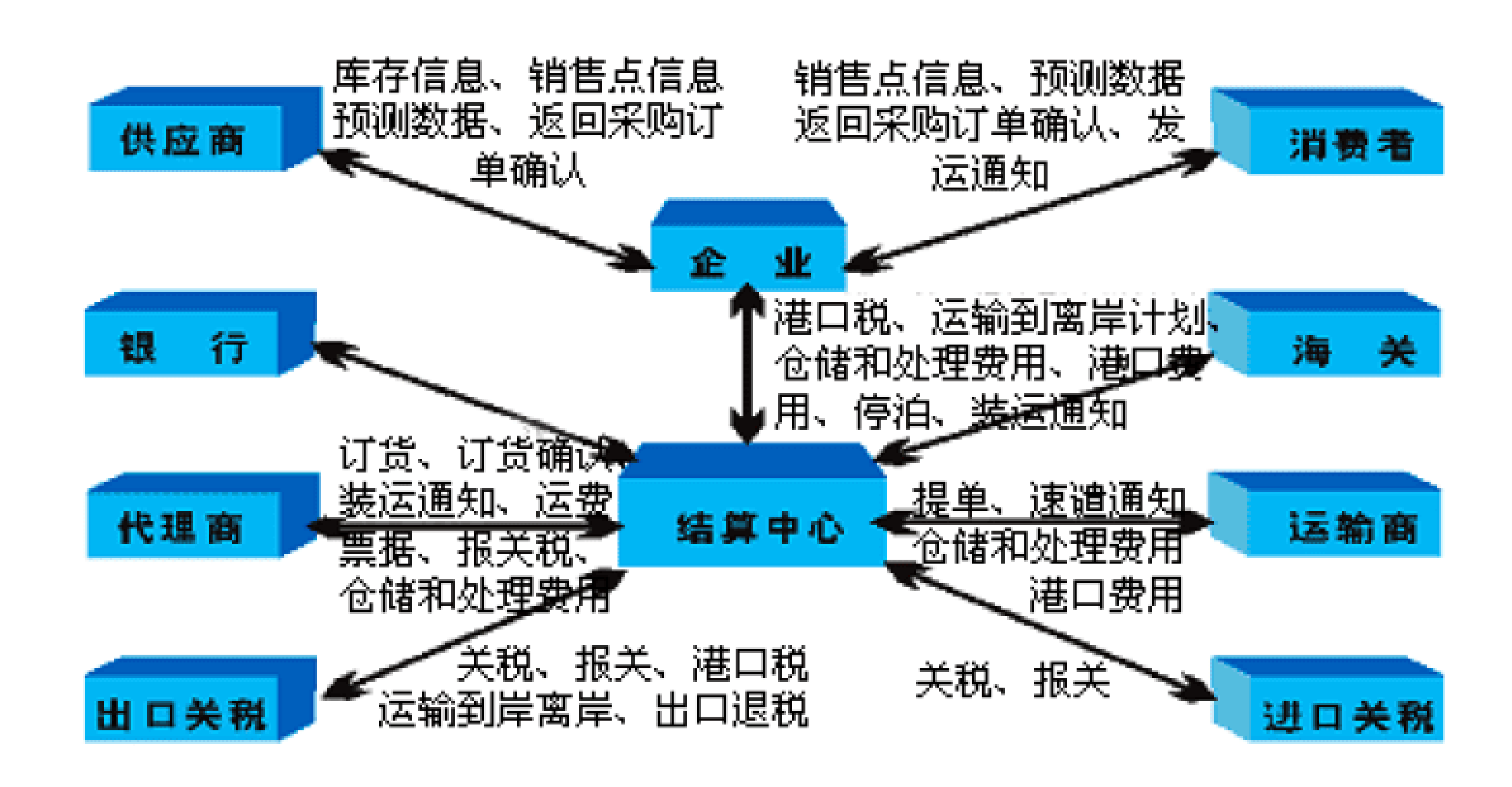


图 3-20　基于 EDI 的供应链信息组织与集成模式

沃尔玛全球电子商务总监 Stephen O’Sullivan 称，沃尔玛实验室计划将沃尔玛的 10 个不同的网站整合成一个，同时将一个 10 个节点的 Hadoop 集群扩展到 250 个节点的 Hadoop 集群。目前，实验室正在设计几个能将当前像 Oracle、Neteeza 这样的开放资源的数据库进行迁移、整合的工具。

沃尔玛还通过先进的大数据预测分析技术发现两个电子产品连锁店 Source 和 Carlie Brown 的顾客的购买意向正在向高档产品转移，并及时调整了两家店的库存，一举将销售业绩提升了 40%。大数据分析技术使得沃尔玛能够实时对市场动态做出积极响应。通过对消费者的购物行为等非结构化数据进行分析，沃尔玛成为最了解顾客购物习惯的零售商，并创造了“啤酒与尿布”的经典商业案例。

沃尔玛曾进行了一系列的收购，包括 Kosmix（沃尔玛实验室前身）、Small Society、

Set Direction、OneRiot、Social Calenda、Grabble 等多家中小型创业公司，这些创业公司要么精于数据挖掘和各种算法，要么在移动社交领域有其专长，由此可见沃尔玛进军移动互联网和挖掘大数据的决心。

【案例解析】：从沃尔玛投入巨资开发大数据工具并推动大数据技术发展的案例中，笔者发现对大数据最热心的企业不是 IT 厂商，而是能直接从大数据中获益的传统企业，他们已经迫不及待，甚至跑到了厂商的前面。

线下零售的海量数据一旦可以整合，必将极大改变现有商业模式。零售巨头沃尔玛正在变革其电子商务模式，而大数据是这次变革的动因。如今，沃尔玛在大数据上的投资已经开始产生回报。相信在沃尔玛的带领下，传统行业也会慢慢意识到大数据的重要性，加速步入大数据时代。

3.3.3 【案例】Clustrix 挖掘整合海量数据

Clustrix 公司创建于 2005 年，Clustrix 总部设在美国旧金山，研发中心设在西雅图。为打开欧洲市场，公司计划将总部迁至荷兰的阿姆斯特丹，还在印度设立了办公室。2010 年，Clustrix 推出了一个可高度扩容的伸缩式数据库解决方案 Sierra，其提供了和 SQL 数据库相似的功能，同时还能对数据存储进行无限制扩展。

Clustrix Sierra 被业内称之为云计算时代的 MySQL，它可以帮助现在要处理海量数据的公司更快地找到数据并解决日益增长的数据扩容等问题。Clustrix Sierra 可以为 SQL 数据库提供专利数据应用方法，帮助人们处理大量的数据，使 SQL 数据库无限扩容成为可能。

【案例解析】：除了传统的大企业已经开始进入大数据领域之外，还有不少的创业企业也意识到了大数据带来的商机，纷纷推出自己的产品，以期抓住大数据时代的机遇，Clustrix 便是其中之一。

笔者在前面的章节已经介绍过，大数据的容量往往是 PB 级别，甚至有些用户的数据量开始达到 EB 级别，这要求未来的存储系统能够具备容量大、易扩展的特点。对海量的、无意义的“非结构化数据”进行挖掘提取，整合成结构化数据，并使之有意义或创造价值，这是很多大数据公司的根本愿望。而完成这些任务有一个前提，必须构架一个大数据分析平台，并利用该平台从海量数据中找到你需要的那部分，这就是创业公司 Clustrix 正在做的。

3.3.4 【案例】长虹联手 IBM 掘金大数据

2013 年 9 月 16 日，IBM 与长虹集团正式发布“绵阳 IBM 大数据分析竞争力中心”。据悉，该中心将以大数据分析和科学管理推动长虹集团智能战略实施和自身转型发展，

从而实现绵阳市智慧城市的落地。

早在 1999 年，长虹就成功使用 ERP 系统对集团进行系统化管理。在家电领域，长虹是最早使用 ERP 系统进行管理的企业。ERP 系统已经成为了长虹信息化的 DNA，也是长虹现阶段发展大数据战略的关键基础。

2008 年，长虹集团成立了虹信公司，开始对外输出软件业务，让更多企业能使用到长虹信息化的成果。

2012 年，长虹虹信公司的收入达到了 2.5 亿元，为中国西南片区的酒类行业、巴斯夫、中海油、云天化等提供了系统的专项服务。

对于长虹来说，大数据服务并不是新起楼阁，随着长虹家庭互联网技术的成熟与整套产品的落地，云计算、大数据服务这些新兴业务将成为公司新的增长领域。而围绕大数据商业模式创新的长虹已积极展开多项相关技术合作开发，包括与中科院软件所进行大数据的数据挖掘项目合作，与中科大进行数据存储、图像识别、算法、云服务平台关键技术等方面的合作，与西安交大共同研发人脸识别、手势识别等技术。

例如，中国首款电视操作系统轩辕 TVOS、全球首创的电视语音浏览器、超高清数字电视系统等，这些软件的研发为长虹带来一个更宽阔的视野，从单一智能终端走向多个智能终端的连接、交互、协同，这是对现有智能终端形态的一次大的颠覆。

在大数据智能时代，长虹芯片将是长虹智能产品的“大脑”，而软件（操作系统）将是“思想”，二者缺一不可。装备了长虹智能芯片和软件的第三代智能电视可以产生很多有趣的应用场景，例如电视节目向不同终端推送，电视控制调节冰箱、空调的状态，以及基于共同的内容产生的社交圈子等。

【案例解析】：在本案例中，处于大数据时代的长虹，无论是在硬件还是软件方面，都占据着相当有利的优势，同时更具备了各软件之间的融合以及硬件与软件间的融合，是最有能力把软件和硬件优势进行有效、完美整合的企业。

大数据对于长虹争夺家庭互联网入口的意义在于：它能使长虹的智能电视更“懂”用户，它能帮助用户实现这样一个梦想，“当你坐在沙发上，电视机就会自动打开，并且调到你最喜欢看的频道”。

3.3.5 【案例】LSI 积极创新数据中心变革

LSI 公司（LSI Corporation）是一家总部位于加利福尼亚州米尔皮塔斯（Milpitas）的半导体和软件领先供应商，其为加速数据存储中心与移动网络性能提供了许多领先的解决方案。

近日，LSI 对其数据中心进行了以下两大创新：

- 为了解决闪存错误率高的现象，LSI 创新了新技术 LSI SHIELD。这是一种高级的纠错方法，即便同时使用出错率较高的廉价闪存存储器也能实现企业级的 SSD 耐

久度和数据完整性。

- 针对典型数据库应用，通过 LSI DVC（DuraWrite Virtual Capacity，一种全新的数据压缩技术）功能，其规划出的虚拟容量可以达到原物理容量的三倍。可以理解为新增的虚拟容量可以显著降低每 GB 的用户存储成本。

通过对数据的采集、存储和分析三个领域的深入研究，LSI 不断解决用户在大数据方面的技术难点。

【案例解析】：不可否认我们已经身处大数据洪流中，无时无刻地体验着大数据带来的价值。面对大数据洪流，数据中心的变革已经迫在眉睫，数据中心的基石 IT 基础架构也需要转变。

面对大数据“多元、高速、海量”三个特点，以及未来基础设施足够的规模及经济性，这些因素推动移动计算的架构向数据流架构的转换。为了顺应这种变化，本案例中的 LSI 必须有智能的芯片解决方案，例如闪存、可共享的 DAS 架构以及异构的多核处理器，为迈进全新的数据中心时代做好全面的准备。

4

掌握：数据管理与挖掘

学前提示

在大数据的带动下，企业对于数据分析与检索软件，以及企业数据管理软件的需求将会逐渐增温，并需要专门设计的硬件和软件工具来处理这些大数据。本章主要介绍大数据管理系统、数据挖掘技术和流程，以及相应的应用案例。

要点展示

- 管理数据，解析开源框架 Hadoop
- 挖掘数据，大数据如何去粗存精
- 大数据管理与挖掘应用案例

4.1 管理数据，解析开源框架 Hadoop

Hadoop 是一种分析技术，也称“大数据”技术，其可快速收集、传播和分析海量数据。目前，该技术已被广泛用于 Google、Yahoo、Facebook、eBay、LinkedIn、Zynga 等网络服务。

4.1.1 Hadoop 的主要特点

Hadoop 是一个由 Apache 基金会开发的分布式系统基础架构，用户可以在不了解分布式底层细节的情况下，使用它来开发分布式程序，并充分利用集群的威力进行高速运算和存储。简而言之，Hadoop 就是一个可以更容易开发和运行处理大规模数据的软件 平台。

Hadoop 的主要特点如下：

- 可靠性（Reliable）。Hadoop 能自动地维护数据的多份备份，并且在任务失败后能自动地重新部署（redeploy）计算任务。
- 扩容能力（Scalable）。Hadoop 能可靠地（reliably）存储和处理千兆字节（PB）数据。
- 高效率（Efficient）。通过分发数据，Hadoop 可以在数据所在的节点上并行地（parallel）处理它们，这使得处理非常快速。
- 成本低（Economical）。可以通过普通机器组成的服务器群来分发以及处理数据。另外，这些服务器群总计可达数千个节点。

专家提醒

Hadoop Distributed File System，简称 HDFS，是一个分布式文件系统。HDFS 有着高容错性（fault-tolerent）的特点，并且设计用来部署在低廉的（low-cost）硬件上。而且它提供高传输率（high throughput）来访问应用程序的数据，适合那些有着超大数据集（large data set）的应用程序。HDFS 放宽了（relax）POSIX 的要求（requirements），这样可以流的形式访问（streaming access）文件系统中的数据。

4.1.2 Hadoop 的发展历史

Hadoop 的源头是 Apache Nutch，该项目始于 2002 年，是 Apache Lucene 的子项目之一。Lucene 是一个功能全面的文本索引和查询库，开发者可以使用 Lucene 引擎方便地在文档上添加搜索功能。例如，桌面搜索、企业搜索以及许多领域特定的搜索引擎

使用的都是 Lucene。

Lucene、Nutch 和 Hadoop 这 3 个项目都是由 Doug Cutting 所创立的，每个项目在逻辑上都是前一个项目的演进。Doug Cutting 起初的目标是从头开始构建一个网络搜索引擎，这样不但要编写一个复杂的、能够抓取和索引网站的软件，还需要面临没有专有运行团队支持运行它的挑战，因为它有很多的独立部件。Doug Cutting 意识到，他们的架构将无法扩展到拥有数十亿网页的网络。

在 2004 年左右，Google 发表了两篇论文来论述 Google 文件系统（GFS）和 MapReduce 框架。Google 声称使用了这两项技术来扩展自己的搜索系统。具体而言，GFS 会省掉管理所花的时间，如管理存储节点。

Doug Cutting 立即看到了这些技术可以适用于 Nutch，接着他的团队实现了一个新的框架，将 Nutch 移植上去，即 Nutch 的分布式文件系统（NDFS）。这种新的技术马上提升了 Nutch 的可扩展性，它开始能够处理几亿个网页，并能够运行在几十个节点的集群上。Doug Cutting 认识到设计一个专门的项目可以充实两种网络扩展所需的技术，于是就有了 Hadoop。

2006 年 1 月，Doug Cutting 加入雅虎（Yahoo），雅虎为他提供一个专门的团队和资源，准备将 Hadoop 发展成一个可在网络上运行的系统。两年后，Hadoop 成为 Apache 的顶级项目。

2008 年 2 月，雅虎宣布其索引网页的生产系统采用了在 10000 多个核的 Linux 集群上运行的 Hadoop。此时，Hadoop 才真正达到了万维网的规模。通过这次机会，Hadoop 成功地被雅虎之外的很多公司应用，如 Last.fm、Facebook 和《纽约时报》。

Hadoop 这个名字不是一个缩写，它是一个虚构的名字。为软件项目命名时，Doug Cutting 似乎总会得到家人的启发。Lucene 是他妻子的中间名，也是她外祖母的名字。他的儿子在咿呀学语时，总把所有用于吃饭的词叫成 Nutch，后来儿子又把一个黄色大象毛绒玩具叫做 Hadoop。Doug Cutting 说：“我的命名标准就是简短，容易发音和拼写，没有太多的意义，并且不会被用于别处。所以，我尝试生活中以前没有人用过的各种词汇，而孩子们很擅长创造单词。”

4.1.3 Hadoop 的主要用途

得益于市场的宣传，企业用户对于“大数据”这一概念的接受程度越来越高，作为一个较为廉价并且开源的大数据解决方案——Hadoop，也越来越受到用户的关注。

那么，选用 Hadoop 系统能够为我们带来什么作用呢？

首先，Hadoop 的方便和简单让其在编写和运行大型分布式程序方面占尽优势。Hadoop 采用分布式存储方式来提高数据读写速度和扩大存储容量；采用 MapReduce 整合分布式文件系统上的数据，保证高速分析处理数据；与此同时还采用存储冗余数据

来保证数据的安全性。

即使是在校的大学生也可以快速、廉价地建立自己的 Hadoop 集群。另一方面，它的健壮性和可扩展性又使它胜任雅虎和 Facebook 最严苛的工作。这些特性使 Hadoop 在学术界和工业界都大受欢迎。如图 4-1 所示为 Hadoop 的主要用途。

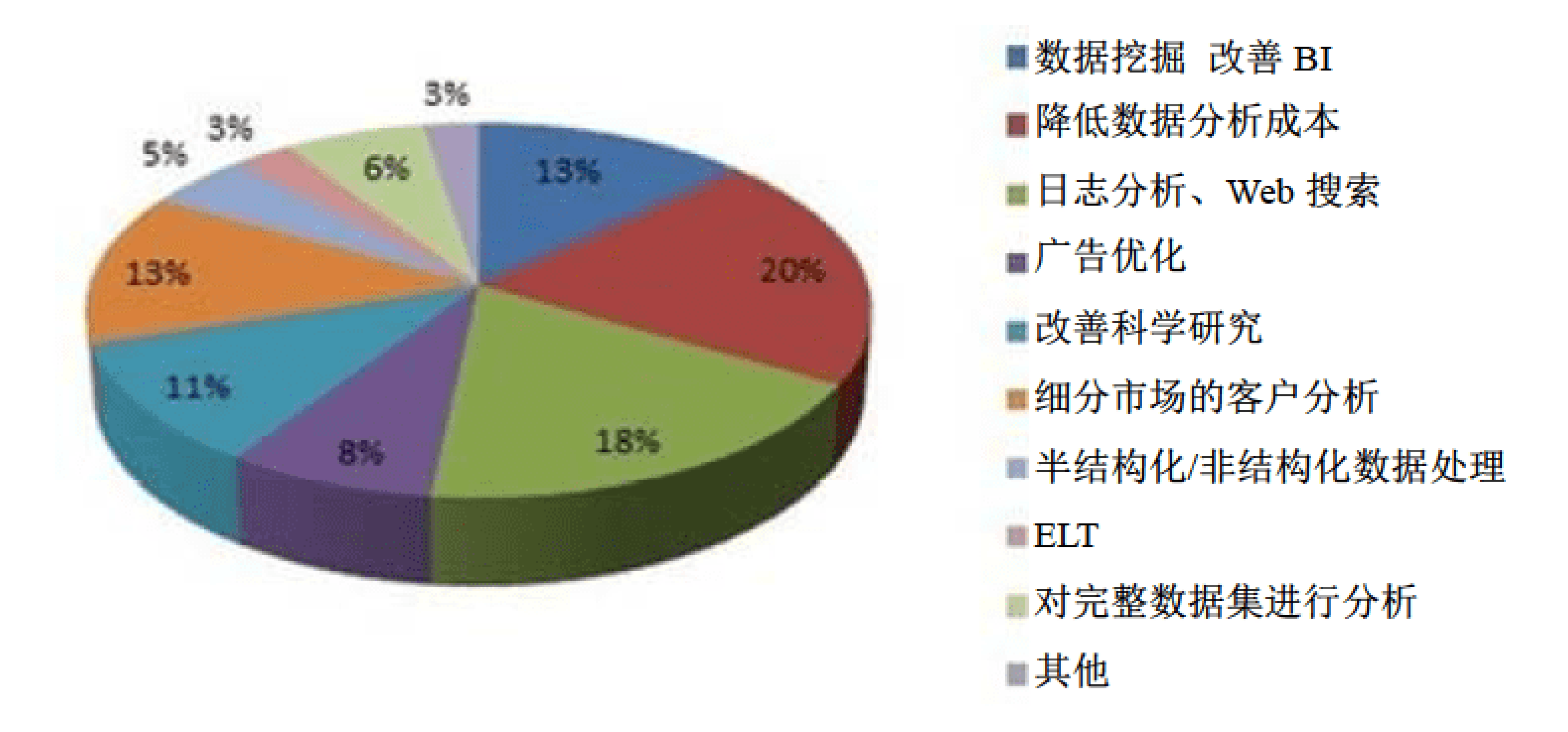


图 4-1 Hadoop 的主要用途

专 家 提 醒

Hadoop 中的 HDFS 具有高容错性，并且是基于 Java 语言开发的，这使得 Hadoop 可以部署在低廉的计算机集群中，同时不限于某个操作系统。Hadoop 中 HDFS 的数据管理能力、MapReduce 处理任务时的高效率以及它的开源特性，使其在同类分布式系统中大放异彩，并在众多行业和科研领域中被广泛应用。

4.1.4 Hadoop 的项目结构

近几年分布式系统的发展越来越快，而 Hadoop 整套项目也起到了推波助澜的作用，而且 Hadoop 已经发展成为包含很多项目的集合。

Hadoop 项目包括 3 部分：Hadoop Distributed File System（HDFS，分布式文件系统）、Hadoop MapReduce 模型和 Hadoop Common。虽然其核心内容是 MapReduce 和 Hadoop 分布式文件系统，但与 Hadoop 相关的 Common、Avro、Chukwa、Hive、HBase 等项目也是不可或缺的。它们提供了互补性服务或在核心层上提供了更高层的服务。Hadoop 的项目结构如图 4-2 所示。

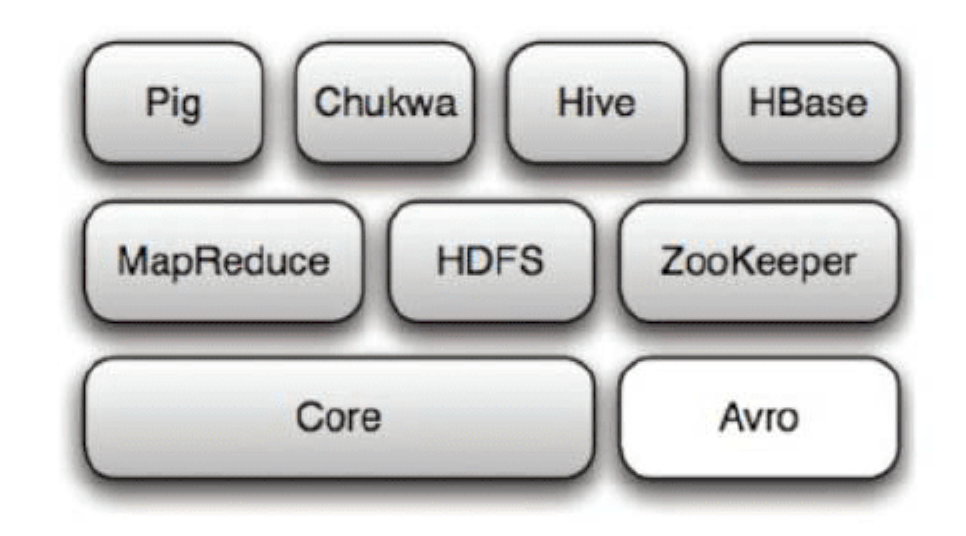


图 4-2 Hadoop 的项目结构

下面对 Hadoop 的各个关联项目进行更详细的介绍。

➢ Pig：一种编程语言，它简化了 Hadoop 常见的工作任务。Pig 可加载数据、表达

转换数据以及存储最终结果。

- Chukwa：Chukwa 是一个开源的用于监控大型分布式系统的数据收集系统，其可以用于监控大规模（2000+以上的节点，每天产生数据量在 TB 级别）Hadoop 集群的整体运行情况并对它们的日志进行分析。Chukwa 是构建在 Hadoop 的 HDFS 和 MapReduce 框架之上的，继承了 Hadoop 的可伸缩性和鲁棒性。Chukwa 还包含了一个强大和灵活的工具集，可用于展示、监控和分析已收集的数据。
- Hive：Hive 是基于 Hadoop 的一个数据仓库工具，可以将结构化的数据文件映射为一张数据库表，并提供简单的 SQL 查询功能，其可以将 SQL 语句转换为 Map Reduce 任务进行运行。Hive 的优点是学习成本低，可以通过类 SQL 语句快速实现简单的 MapReduce 统计，不必开发专门的 MapReduce 应用，十分适合数据仓库的统计　分析。
- HBase：HBase 是一个分布式的、面向列的开源数据库，类似 Google BigTable 的分布式 NoSQL 列数据库。HBase 不同于一般的关系数据库，它是一个适合于非结构化数据存储的数据库。
- MapReduce：MapReduce 是一种编程模型，用于大规模数据集（大于 1TB）的并行运算。MapReduce 极大地方便了编程人员的工作，即使在不了解分布式并行编程的情况下，也可以将自己的程序运行在分布式系统上。MapReduce 在执行时先指定一个 Map（映射）函数，其把输入键值对映射成一组新的键值对，经过一定处理后交给 Reduce（化简），Reduce 对相同 key 下的所有 value 进行处理后再输出键值对作为最终的结果。
- HDFS：HDFS 是一个分布式文件系统。HDFS 原本是开源的 Apache 项目 Nutch 的基础结构，最后它却成为了 Hadoop 基础架构之一。HDFS 放宽了对可移植操作系统接口（Portable Operating System Interface，POSIX）的要求，这样可以实现以流的形式访问文件系统中的数据。
- ZooKeeper：ZooKeeper 是一个针对大型分布式系统的可靠协调系统，提供的功能有配置维护、名字服务、分布式同步、组服务等。ZooKeeper 的目标就是封装好复杂易出错的关键服务，将简单易用的接口和性能高效、功能稳定的系统提供给用户，提供类似 Google Chubby（分布式锁服务）的功能。
- Core（酷睿）：酷睿是一款由英特尔设计的节能新型微架构，设计的出发点是提供卓然出众的性能和能效，提高每瓦特性能，也就是所谓的能效比。
- Avro：Avro 是用于数据序列化的系统，其提供了丰富的数据结构类型、快速可压缩的二进制数据格式、存储持久性数据的文件集、远程调用 RPC 的功能和简单的动态语言集成功能。其中代码生成器既不需要读写文件数据，也不需要使用或实现 RPC 协议，它只是一个可选的对静态类型语言的实现。

专 家 提 醒

Common 是为 Hadoop 其他子项目提供支持的常用工具，它主要包括 FileSystem、RPC 和串行化库。它们为在廉价硬件上搭建云计算环境提供基本的服务，并且会为运行在该平台上的软件开发提供所需的 API。在 Hadoop 0.20 及以前的版本中，包含 HDFS、MapReduce 和其他项目公共内容，从 0.21 开始 HDFS 和 MapReduce 被分离为独立的子项目，其余内容为 Hadoop Common。

4.1.5 Hadoop 的体系结构

Hadoop 的整个体系结构主要是通过 HDFS 来实现对分布式存储的底层支持，并且通过 MapReduce 来实现对分布式并行任务处理的程序支持。可以说，HDFS 和 MapReduce 是 Hadoop 的两大核心体系结构。

1. HDFS 的体系结构

HDFS 是一个主从结构（Master/Slave）模型，一个 HDFS 集群是由一个 NameNode 和若干个 DataNode 组成的。如图 4-3 所示为 HDFS 的体系结构。

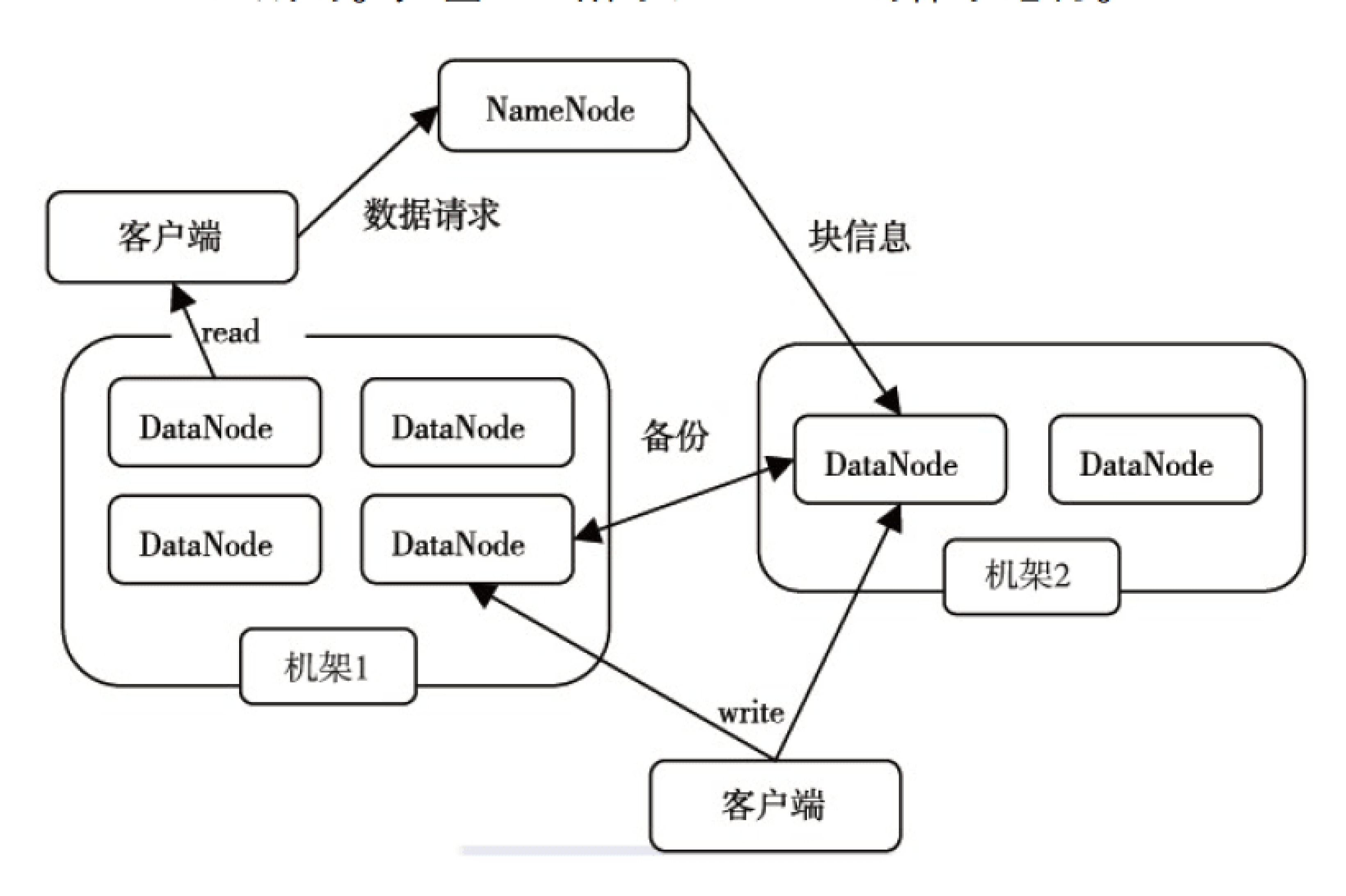


图 4-3 HDFS 的体系结构

- NameNode（名称节点）：NameNode 作为主服务器，管理文件系统的命名空间和客户端对文件的访问操作。
- DataNode（数据节点）：集群中的 DataNode 管理存储的数据。HDFS 允许用户以文件的形式存储数据，从内部来看，文件被分成若干个数据块，而且这若干个数据块存放在一组 DataNode 上。

NameNode 执行文件系统的命名空间操作，例如，打开、关闭、重命名文件或目录

等，它也负责数据块到具体 DataNode 的映射。DataNode 负责处理文件系统客户端的文件读写请求，并在 NameNode 的统一调度下进行数据块的创建、删除和复制工作。

2. MapReduce 的体系结构

MapReduce 是一种并行编程模式，这种模式使得软件开发者可以轻松地编写出分布式并行程序。MapReduce 框架是由一个单独运行在主节点上的 Job Tracker 和运行在每个集群从节点上的 Task Tracker 共同组成的。当一个 Job 被提交时，Job Tracker 接收到提交作业和其配置信息之后，就会将配置信息等分发给从节点，同时调度任务并监控 Task Tracker 的执行。

很多人也许看不明白，下面笔者举个简单的例子来说明 MapReduce 结构的作用。假设你是幼儿园的老师，带着一群小朋友做一个加减乘除的游戏，你给每一个小朋友出一道题目，然后让他算好后给你报告答案，你再给他出一道题目，周而复始如此做。如果只有十几个小朋友在算，相信你可以轻松应付；如果上了一百个小朋友，估计每个人都会争着表现，叫嚷着让你出题，这时你肯定会感到不堪重负。

面对这样的场景，我们通常的经验是“再拽的算法也难以抵挡海量的数据或任务”。因此，应对方法主要还是增加资源，其次才是优化算法，而且两者可并行。即小朋友在增加的同时，我们也相应地增加老师的数量，通过这样的途径来缓解每个老师的压力。

与这种场景类似，MapReduce 结构也面临类似的问题。越来越多的 Task Tracker（小朋友）会让有限的 Job Tracker（老师）很有压力，以至于 Task Tracker 有很多时，Job Tracker 不能及时响应请求，很多 Task Tracker 就让资源空闲着，等待 Job Tracker 的 response（响应）。因此，如何优化 MapReduce 结构，也是各个大数据分析平台急需解决的难题。

总之，HDFS 和 MapReduce 共同组成了 Hadoop 分布式系统体系结构的核心。HDFS 在集群上实现了分布式文件系统，MapReduce 在集群上实现了分布式计算和任务处理。HDFS 在 MapReduce 任务处理过程中提供了对文件操作和存储等的支持，MapReduce 在 HDFS 的基础上实现了任务的分发、跟踪、执行等工作，并收集结果，二者相互作用，完成了 Hadoop 分布式集群的主要任务。

4.2 挖掘数据，大数据如何去粗存精

数据挖掘（Data Mining）是数据库知识发现（Knowledge-Discovery in Databases，KDD）中的一个重要步骤。数据挖掘一般是指从大量的数据中通过算法搜索隐藏于其中信息的过程。数据挖掘是通过分析每个数据，从大量数据中寻找其规律的技术，其一般流程如图 4-4 所示。

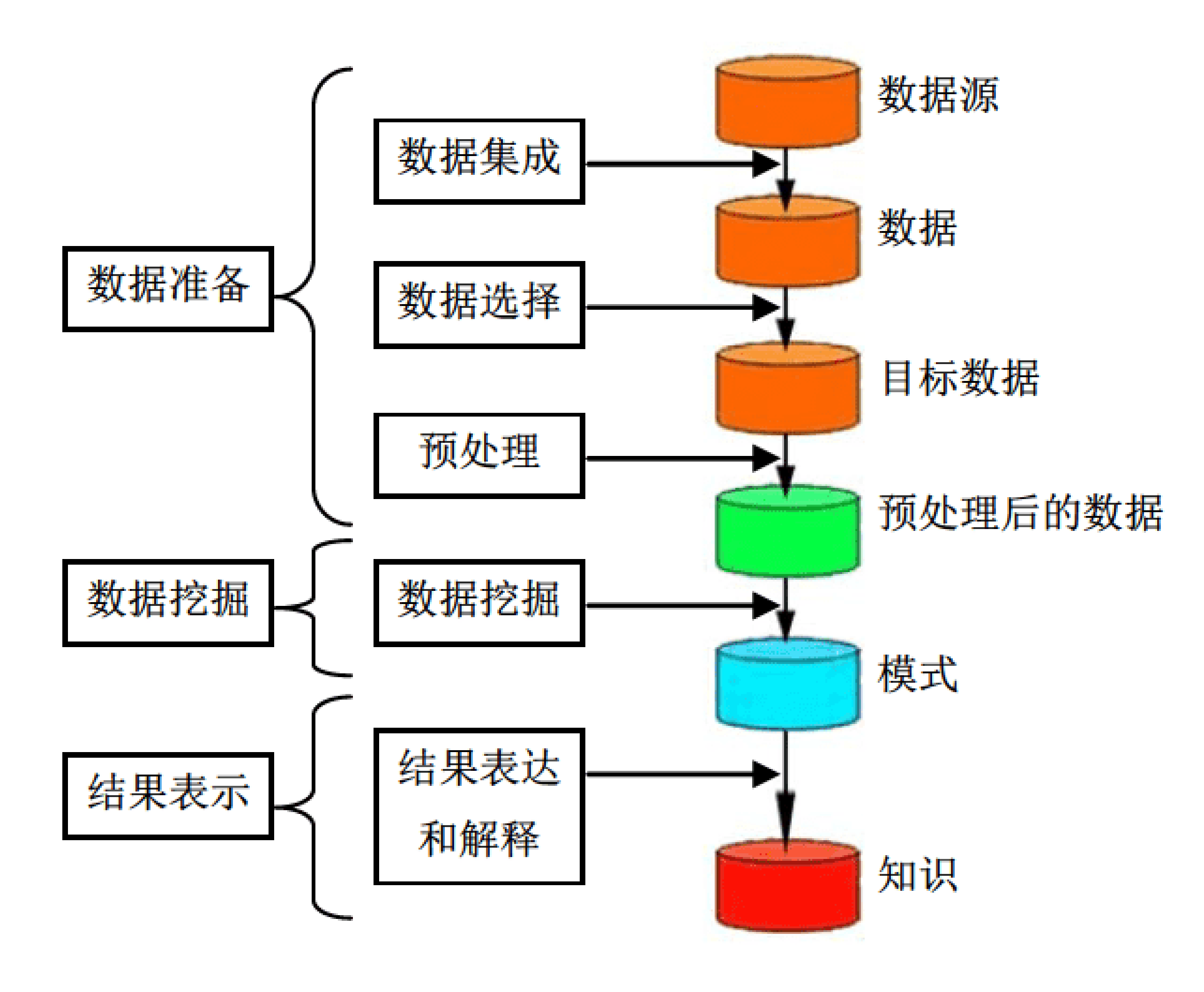


图 4-4 数据挖掘的一般流程

4.2.1 准备数据

数据准备是指从相关的数据源中选取所需的数据并整合成用于数据挖掘的数据集，如表 4-1 所示。

表 4-1 准备数据的流程

准 备 步 骤	具 体 内 容
第一步：选择数据	搜索所有与业务对象有关的内部和外部数据信息，并从中选择出用于数据挖掘的数据
第二步：预处理数据	研究数据的质量，为进一步的分析作准备，并确定将要进行的挖掘操作的类型
第三步：转换数据	将数据转换成分析模型，这个分析模型是针对挖掘算法建立的。建立一个真正适合挖掘算法的分析模型是数据挖掘成功的关键

4.2.2 挖掘过程

数据的挖掘过程是指对所得到的经过转换的数据进行挖掘，其一般流程如表 4-2 所示。

表 4-2 挖掘数据的流程

挖 掘 步 骤	具 体 内 容
第一步：建模（Modeling）	在这个阶段，可以选择和使用不同的模型技术，模型参数被调整到最佳的数值。一般情况下，有些技术可以解决一类相同的数据挖掘问题。有些技术在数据形成上有特殊要求，因此需要经常跳回到数据准备阶段

续表

挖 掘 步 骤	具 体 内 容
第二步：评估（Evaluation）	到项目的这个阶段，你已经从数据分析的角度建立了一个高质量的模型。在最后部署模型之前，重要的事情是彻底地评估模型，检查构造模型的步骤，确保模型可以完成业务目标。这个阶段的关键是确定是否有重要业务问题没有被充分地考虑。在这个阶段结束后，必须达成一个使用数据挖掘结果的决定
第三步：部署（Deployment）	通常，模型的创建不是项目的结束。模型的作用是从数据中找到知识，获得的知识需要重新组织和展现，便于用户使用。根据需求，这个阶段可以产生简单的报告，或是实现一个比较复杂的、可重复的数据挖掘过程。在很多案例中，这个阶段是由客户而不是数据分析人员承担部署的工作

专 家 提 醒

在客户生命周期的过程中，各个不同的阶段包含了许多重要的事件。数据挖掘技术可以应用于客户生命周期的各个阶段以提高企业客户关系管理能力，包括争取新的客户，让已有的客户创造更多的利润，保持住有价值的客户等。

4.2.3 结果表示

结果表示是指根据客户的决策要求，对挖掘出的信息进行分析，抽取出最有价值的部分，通过决策支持工具提交给决策者。结果表示的一般流程如表 4-3 所示。

表 4-3 结果表示的流程

挖 掘 步 骤	具 体 内 容
第一步：结果分析	解释并评估结果，其使用的分析方法一般应视数据挖掘操作而定，通常会用到可视化技术，如图 4-5 所示
第二步：知识的同化	将分析所得到的知识集成到业务信息系统的组织结构中去

在数据挖掘中发现的知识可直接用于指导 OLAP 的分析处理，而 OLAP 分析得到的新知识也可以直接补充到系统的知识库中。为增强数据挖掘的效率，可以将粗糙集理论与神经网络、遗传算法、模糊数学、决策树等方法相结合。一般情况下，粗糙集理论用于产生确定规则，神经网络用于产生非确定规则，粗糙集理论的使用提高了系统的运算速度，同时神经网络则使产生的规则集泛化能力提高。

大数据是一种具有隐藏法则的“人造自然系统”，寻找大数据的科学模式将带来对研究大数据之美的一般性方法的探究，尽管这样的探索十分困难，但是如果我们找到了

将非结构化、半结构化数据转化成结构化数据的方法，已知的数据挖掘方法将成为大数据挖掘的工具。

要表达的数据和信息	建议采用图形					
	饼图	垂直柱	水平柱	线图	水泡	其他
整体的一部分						
不同数据的比较						
时间序列						
频率						
两组数据的相关性						
和多重数据、标准相比较						

图 4-5　可视化的展现

专 家 提 醒

随着企业信息化水平的不断提高，采用基于数据仓库的决策支持系统，能增强管理者的决策能力，获取更好的管理效果和企业竞争优势。

4.3　大数据管理与挖掘应用案例

数据挖掘对企业来说究竟有何意义，不妨先看看以下几个事件。

- 事件 1：当你在网上搜索一条飞往北京的航班信息时，同时看到网站上出现了北京宾馆的打折信息。
- 事件 2：你正在观赏的一部电影，采用了以几十万 GB 数据为基础的计算机图形图像技术。
- 事件 3：被经常光顾的商店在对顾客行为进行数据挖掘的基础上可获取最大化的利润。
- 事件 4：用算法预测人们购票需求，航空公司以不可预知的方式调整价格。
- 事件 5：智能手机的 APP 应用识别到你的位置，因此你会收到附近餐厅的服务信息。

这些都是对海量数据进行挖掘分析的结果。笔者觉得一个数据库只要有几十万条以

上的记录，就有数据挖掘的价值。本节主要介绍大数据管理与挖掘的应用案例，希望对你有一定的启发和学习价值。

4.3.1 【案例】用数据挖掘筛查高危病人

通常情况下，医生会通过一系列检查来确定人们的健康情况。然而，麻省理工学院的研究者约翰·古塔格（John Guttag）和柯林·斯塔尔兹（Collin Stultz）创建了一个计算机模型来分析心脏病患者被弃用的心电图数据，如图 4-6 所示。



图 4-6 约翰·古塔格（John Guttag）和柯林·斯塔尔兹（Collin Stultz）

他们利用数据挖掘和机器学习的方法在海量的数据中筛选，发现心电图中出现三类异常者一年内死于第二次心脏病发作的几率比未出现者高一至二倍。这种新方法能够识别出更多的、无法通过现有的风险筛查技术来探查的高危病人。

【案例解析】：如何应对“大数据”，是摆在医院 IT 部门面前的一个“大考验”。如果处理不好，“大数据”就会成为“大包袱”、“大问题”；反之，如果应用得当，“大数据”则会为医院带来“大价值”。而这一切，都离不开科学地规划和部署存储架构。

当每个老百姓都可以随时管理、查询自己的健康医疗数据时，而且这样的数据将不局限于体检结果、就诊记录，还可以衍生到你的基因数据，你的日常健康行为监测数据，医疗大数据的价值才能真正发挥，人类对自身的认识也将上一个新的台阶。

4.3.2 【案例】数据挖掘助力 NBA 赛事

美国著名的国家篮球队 NBA 的教练，利用 IBM 公司提供的数据挖掘工具临场决定替换队员。想象你是 NBA 的教练，你靠什么带领你的球队取得胜利呢？当然，最容易

想到的是全场紧逼、交叉扯动和快速抢断等具体的战术和技术。

如今，数据挖掘成了 NBA 教练们的新式武器。据悉，大约 20 个 NBA 球队使用了 IBM 公司开发的数据挖掘应用软件 Advanced Scout 来优化他们的战术组合。Advanced Scout 是一个数据分析工具，教练可以用便携式电脑在家里或在路上挖掘存储在 NBA 中心的服务器上的数据。每一场比赛的事件都被统计分类，如得分、助攻、失误等。因为有时间标记，教练可非常容易地通过搜索 NBA 比赛的录像来理解统计发现的含义。

例如，魔术队利用 Advanced Scout 系统分析显示：先发阵容中的两个后卫安佛尼・哈德卫（Anfernee Hardaway）和伯兰・绍（Brian Shaw）在前两场中被评为-17 分，这意味着他俩在场上，本队输掉的分数比得到的分数多 17 分。然而，当哈德卫与替补后卫达利尔・阿姆斯创（Darrell Armstrong）组合时，魔术队得分为+14 分。因此，魔术队在下一场比赛中特意增加了阿姆斯创的上场时间。

结果显而易见，阿姆斯创得了 21 分，哈德卫得了 42 分，魔术队以 88 比 79 获胜。因此，魔术队在第四场继续让阿姆斯创先发，再一次打败了热队。在第五场比赛中，这个靠数据挖掘支持的阵容没能拖住热队，但 Advanced Scout 毕竟帮助了魔术队赢得了打满 5 场，直到最后才决出胜负的机会。

另外，教练们通过 Advanced Scout 系统，可以在对方球员与自己的队员在“头碰头”的瞬间分解双方接触的动作，进而设计合理的防守策略。

【案例解析】：Advanced Scout 的开发人员布罕德瑞表示：“教练们可以完全没有统计学的培训经历，但他们可以利用数据挖掘制定策略”。开发者还可以继续开发出与 Advanced Scout 相似的数据挖掘应用，增加其功能，可以让教练、广播员、新闻记者及球迷挖掘其他数据统计。

专家提醒

需要注意的是，所有电脑系统都有其局限性，因此你不要期望这样的数据挖掘可以帮助一支球队找到赢得足球世界杯的策略。

4.3.3 【案例】用数据挖掘控制鲜花库存

Pro Flowers 是美国著名的鲜花在线预订网站，有四万多家连锁花店提供配送服务。其网站也制作得相当精美，不同主题的鲜花图片非常地赏心悦目，如图 4-7 所示。

由于鲜花极易枯萎，Pro Flowers 不得不均匀地削减库存，否则可能导致一种商品过快售罄或库存鲜花濒于凋谢。

另外，由于日交易量较高，Pro Flowers 的网站管理人员需要对零售情况进行大量的分析，例如，转换率，也就是多少页面浏览量将导致销售产生。例如，如果 100 人中仅有 5 人看到玫瑰时就会购买，而盆景的转换率则为 100 比 20，那么不是页面设计有

问题，就是玫瑰的价格有问题。此时，Pro Flowers 就要迅速对网站上的玫瑰价格进行调整。对于可能过快售罄的商品，Pro Flowers 通常不得不在网页中弱化该商品或取消优惠价格，从而设法减缓该商品的销售。



图 4-7　Pro Flowers 网站主页

过去，这一工作通常由人工来完成，效率极其低下。Pro Flowers 营销副总裁 Chris d'Eon 表示："自己分析数据是浪费时间。我们需要一种浏览数据的方式，能够让我们即刻采取行动。"

因此，Pro Flowers 采用了 WebSideStory 推出的数据挖掘 ASP 服务——HitBox，其可以使企业的计划者在业务高峰日也能够对销售情况做出迅速反应。WebSideStory 为 700 多家公司提供多种在线访客页面点击的跟踪服务，每月为公司分析超过 300 亿个网页。采用 HitBox 后，Pro Flowers 的网站管理人员可以借助便于阅读的可视化界面来了解销售数据和转换率，节省了工作效率。

HitBox 是分析领域的新突破，它将 WebSideStory 专业的、实时的数据收集体系架构与挖掘数据的能力整合在一起，结果得到快速反应的、精确到秒的访问效果，使业务人员大幅提高了在线活动的能力。

作为一种完全托管的 on-demand 服务，HitBox 可实时收集访问者或客户的行为信息，并通过简便的 Web 浏览器界面提供定制数据，这种服务不需要软硬件投资，可以在数天内实施。

【案例解析】：对于商业型企业来说，通过收集、加工和处理涉及消费者消费行为的大量信息，确定特定消费群体或个体的兴趣、消费习惯、消费倾向和消费需求，进而推断出相应消费群体或个体下一步的消费行为，然后以此为基础，对所识别出来的消费群体进行特定内容的定向营销，这与传统的不区分消费者对象特征的大规模营销手段相比，大大节省了营销成本，提高了营销效果，从而为企业带来更多的利润。

4.3.4 【案例】挖掘人类头脑里的大数据

人类连接组项目（Human Connectome Project）是美国国立卫生院 NIH 2009 年开始资助的一个 5 年项目，不同的几个大学/研究所分成两组进行。第一组由圣路易斯华盛顿大学（Washington University in Saint Louis）为首，预计投资 3000 万美元。另一组由哈佛大学、麻省总医院以及 UCLA（University of California，Los Angeles，加利福尼亚大学洛杉矶分校）组成，预计投资 850 万美元。

人类连接组项目旨在通过扫描 1200 名健康成年人的大脑，比较他们大脑各区域神经连接的不同以及如何由此导致认知和行为方面的个体差异，最终描绘出人类大脑的所有神经连接情况。2012 年 12 月 21 日出版的美国《科学》杂志将人类连接组计划列为 2013 年六大值得关注的科学领域之一。

据悉，人类连接组项目使用 3 种磁共振造影观察脑的结构、功能和连接。根据圣路易斯华盛顿大学的连接组项目办事处的信息学主任丹尼尔·马库斯（Daniel Marcus）的预期，数据收集工作完成之时，连接组研究人员将埋首于大约 100 万 GB 数据中工作。一旦绘制出精细的大脑结构、功能图，就可以进一步研究神经环路的构造，大脑随发育、年龄增长的变化，大脑的网络属性，神经/精神疾病的根源；还可以研究出大脑多大程度上由基因决定，以及不同的大脑功能/结构和行为的关系，从而给其他所有的类似研究提供最完美的“金标准”对照。

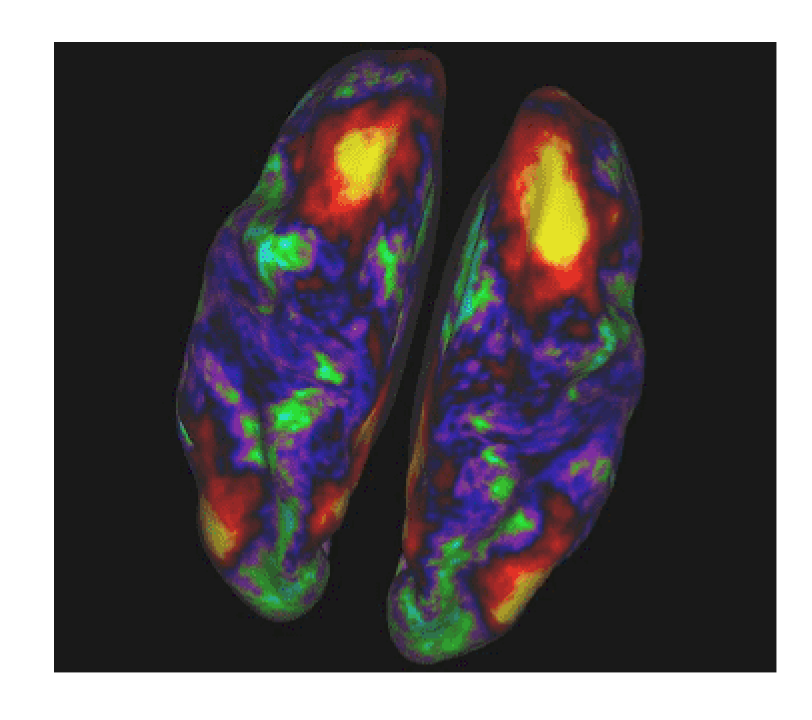


图 4-8　核磁共振扫描出的人类大脑

如图 4-8 所示，为 20 名健康人受试者处于休息状态下接受核磁共振扫描，得到的大脑皮层不同区域间新陈代谢活动的关联关系，并用不同的颜色表现出来。

马库斯说：“我们将拥有 1200 个人的数据，因此我们可以观察到个体之间脑区分布的差别，以及脑区之间是如何关联的。”

专家提醒

除了连接组，人类的身体里面还有很多充满数据的“组”。

- 基因组：由 DNA 编码的全部基因信息，或者由 RNA（核糖核酸）编码的（例如病毒）全部基因信息。
- 转录组：由一个有机体的 DNA 产生的全套 RNA“读数”。
- 蛋白质组：所有可以用基因表达的蛋白质。
- 代谢组：一个有机体在新陈代谢过程中的所有小分子，包括中间产物和最终产物。

【案例解析】: 意识从何而来？思维和智能是如何出现的？这些终极问题都蕴藏在人类的大脑里面。人脑是终极的计算机器，也是终极的大数据困境，因为在独立的神经元之间有无数可能的连接。

人类连接组项目是一项雄心勃勃的试图绘制出不同脑区之间相互作用的计划，是一项对大脑进行的逆向工程研究，目的是充分挖掘大脑里的有效数据，借此明白“大脑”是怎么被建造的，而后就可以再建模拟的“大脑”，从而真正地实现人造智能。

4.3.5 【案例】数据挖掘助力银行的营销

蒙特利尔银行（Bank of Montreal）是根据加拿大《国会法》于 1817 年 11 月 3 日建立的，是加拿大历史最悠久的银行，也是加拿大的第三大银行，至今已有 180 多年的历史。

20 世纪 90 年代中期，行业竞争的加剧导致蒙特利尔银行需要通过“交叉销售”来锁定 1800 万客户。“交叉销售”是指借助 CRM（客户关系管理），发现顾客的多种需求，并通过满足其需求而销售多种相关服务或产品的一种新兴营销方式。

“交叉销售”体现了银行的一个新焦点——客户，而不是商品。银行应该认识到客户需要什么产品以及如何推销这些产品，而不是等待人们来排队购买。然后，银行需要开发相应商品并进行营销活动，从而满足这些需求。

在应用数据挖掘之前，银行的销售代表必须于晚上 6 点至 9 点在特定地区通过电话向客户推销产品。但是，正如每个处于接受端的人所了解的那样，大多数人在工作结束后对于兜售并不感兴趣。因此，在晚餐时间进行电话推销的反馈率非常低。

为了改变这种不利的局面，银行开始采用 IBM DB2 Intelligent Miner Scoring 系统，基于银行账户余额、客户已拥有的银行产品以及所处地点和信贷风险等标准来评价记录档案，这些评价可用于确定客户购买某一具体产品的可能性。另外，该系统能够通过浏览器窗口进行查看，这使得管理人员不必分析基础数据，因此非常适合于非统计专业的人员。

蒙特利尔银行的数据挖掘工具为管理人员提供了大量信息，从而帮助他们对从营销到产品设计的任何事情进行决策。现在，当进行更具针对性的营销活动时，银行能够区别对待不同的客户群，以提升产品和服务质量，同时还能制定适当的价格和设计各种奖励方案，甚至确定利息费用。

【案例解析】: “交叉销售”的核心是向原有顾客销售多种相关的产品和服务，但并不是简单地将顾客还没有购买的本企业的产品和服务推销给顾客，而是通过对顾客数据的分析和应用，发现顾客的不同需求并满足其需求的营销方式。

企业进行“交叉销售”首先要分析现有顾客消费行为的数据，进行顾客赢利性分析（通过顾客细分对顾客进行赢利性分析），使用数据挖掘进行交叉规则的提取并锁定目

标顾客，如图 4-9 所示。

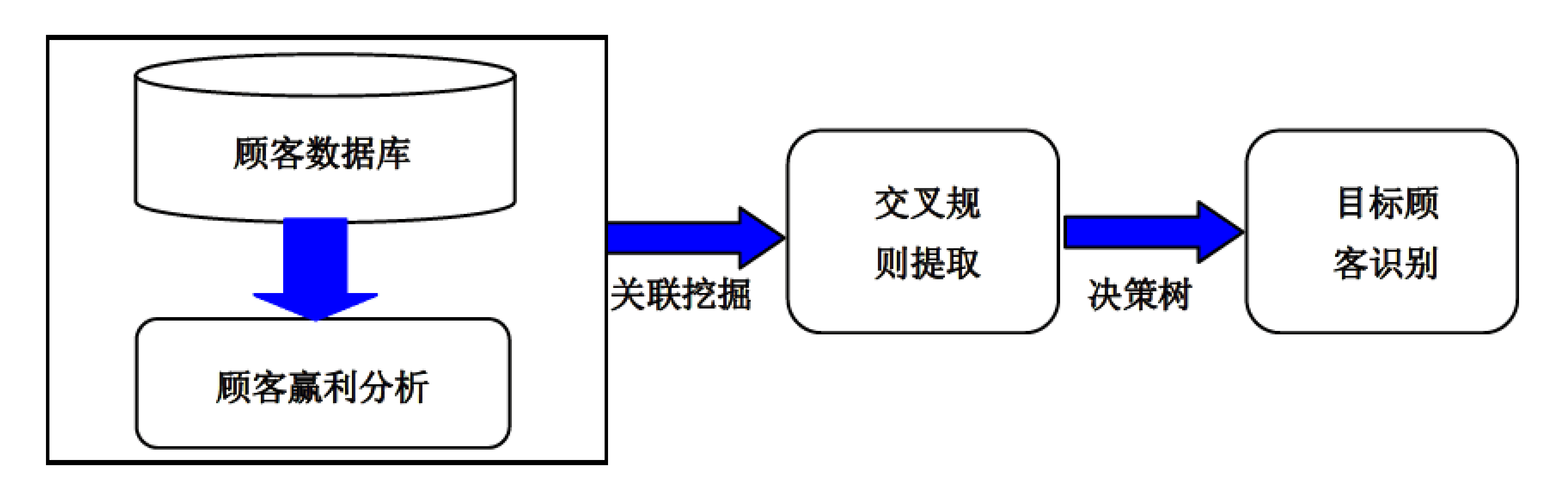


图 4-9 “交叉销售”的数据挖掘过程

专 家 提 醒

数据挖掘技术在企业市场营销中得到了比较普遍的应用，它是以市场营销学的市场细分原理为基础的，其基本假定是“消费者过去的行为是其今后消费倾向的最好说明”。企业通过长期对顾客关系管理工具和数据库的投资，积累了海量顾客数据。对这些数据的深度探索是企业深入了解和掌握现有顾客群的关键，也是实现营销精细化的基础。

4.3.6 【案例】星系动物园里的数据挖掘

星系动物园是英国研究机构开展的天文学研究中一次规模最大的普查活动。志愿者利用网上的图片对 100 万最明亮的“疑似”星系进行识别，分辨出图中究竟是漩涡星系还是椭圆星系，或者根本就不是星系。

星系动物园计划上线 5 年以来，已经有超过 65 万名来自世界各地的天文爱好者参与其中，这些信息帮助科学家发表了多篇高质量的论文。

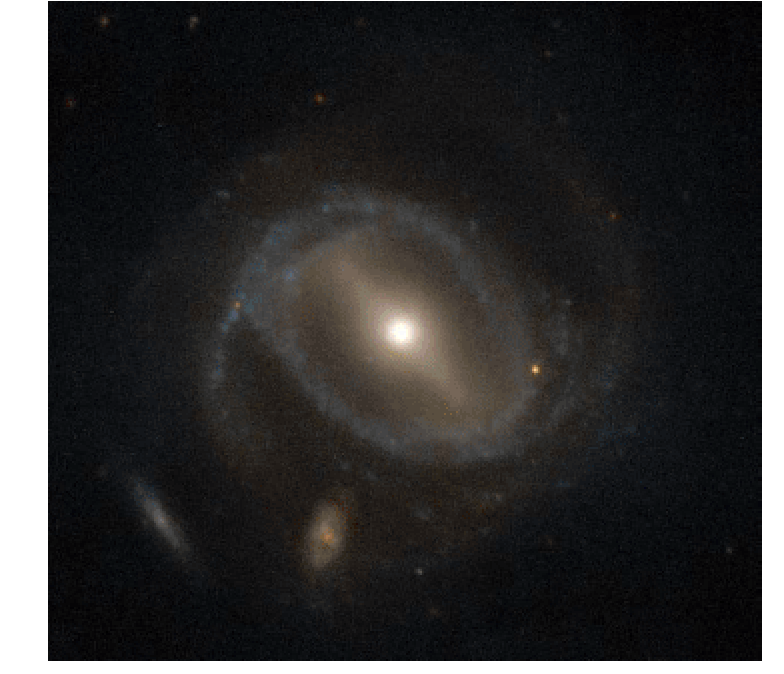


图 4-10 志愿者们发现的背光星系

如图 4-10 所示，为星系动物园的志愿者们发现的差不多 2000 个背光星系之一。它被其后方的另一个星系照亮，来自背后的光令前景星系中的尘埃清晰可辨。星际尘埃在恒星的形成中扮演了关键的角色，但它本身也是由恒星形成的，因此检测其数量和位置对于了解星系的历史至关重要。

下面笔者带你体验一下这个过程：进入 galaxyzoo.org 网站后，注册一个用户名并登录。接受一些简单培训后，就可以在网站上逐个识别照片中的星系，如图 4-11 所示。每个星系照片将由多人反复识别，以减少差错。如果志愿者对某一星系的识别结果不同，天文学家将做出最后判断。

星系动物园积累志愿者们的庞大数据，使之成为计算机学习分类的理想材料，这种动物园方法在 zooniverse.org 网站上得到了复制和优化。zooniverse.org 是一个运行着

大约 20 个项目的机构，这些项目的处理对象包括热带气旋、火星表面和船只航行日志上的气象数据等。

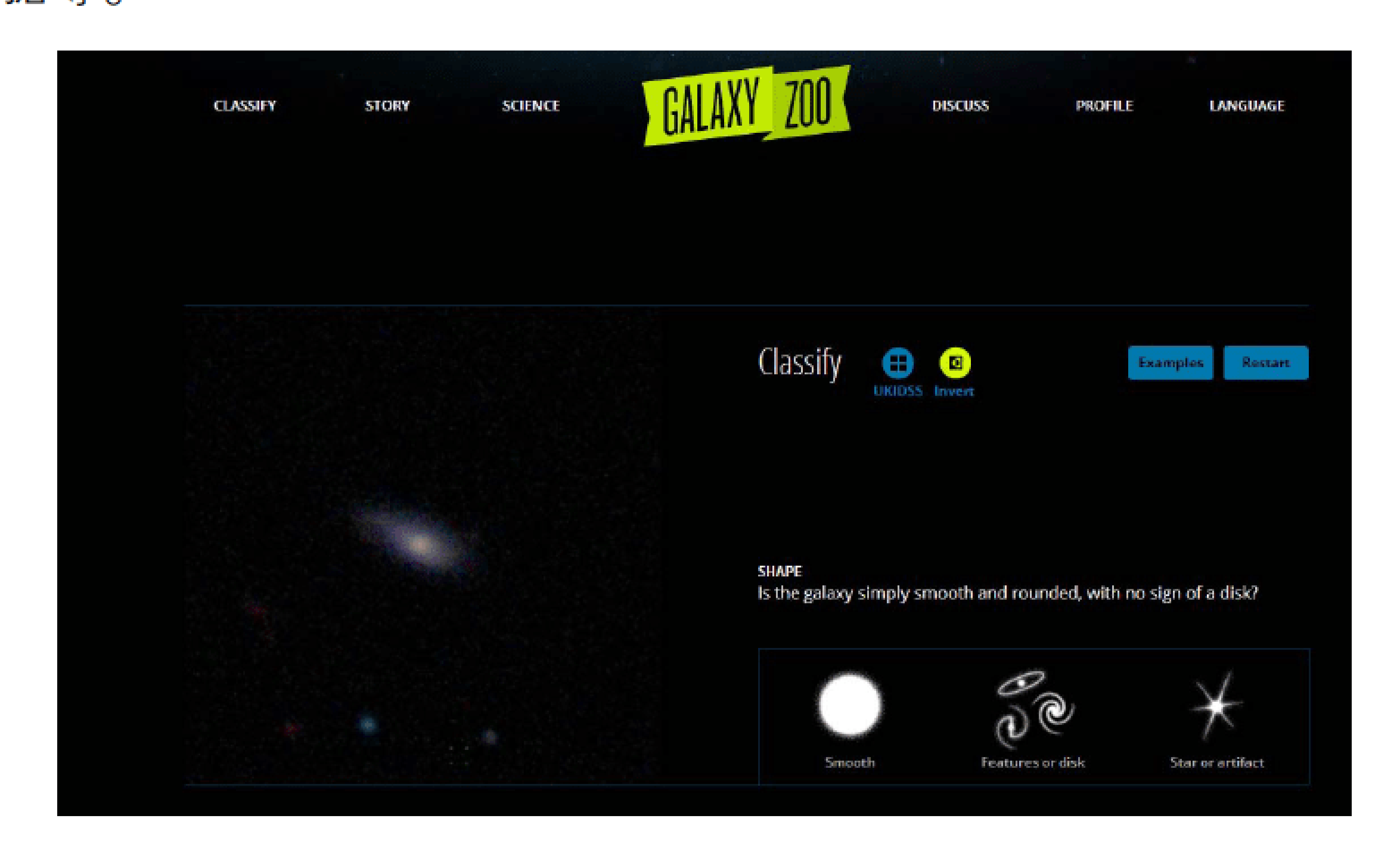


图 4-11　在 galaxyzoo.org 网站上逐个识别照片中的星系

【案例解析】：人脑相比电脑优势在于，合理分类的同时不至于剔除掉那些不规则的、怪异的和令人惊奇的形态。星系动物园项目打破了大数据的规矩：它没有对数据进行大规模的计算机数据挖掘，而是把图像交给活跃的志愿者，由他们对星系做基础性的分类。

星系动物园项目依赖统计学、众多观察者以及处理、检查数据的逻辑。假如观察某个特定星系的人增加时，而认为它是椭圆星系的人数比例保持不变，这个星系就不必再观察了。如果将来中国天文学研究也有海量数据需要挖掘和处理，笔者觉得也可以借鉴这一模式。



读书笔记

5

管理：用数据洞察一切

学前提示

对于大数据，不仅要从数据挖掘、数据分析的层面去解决“大”的问题，更重要的是如何将挖掘与分析的结果直观呈现出来，转换为用户真正需要的有价值的洞察力。本章将结合企业管理和能源管理，释放一切数据的力量，做到真正的智能化管理。

要点展示

- 不能再等，大数据时代的思维变革
- 知己知彼，数据分析的演变与现状
- 企业管理中的大数据分析应用案例
- 能源管理中的大数据分析应用案例

5.1 不能再等，大数据时代的思维变革

“大数据时代”带来了思维模式、商业模式和数据管理控制方式等方面的重大改变，需要我们树立新理念，运用“多平台融合”的信息处理方法，努力对信息进行动态和可视化的呈现。

5.1.1 利用所有的数据

在大数据时代，我们要改变以下 3 个思维：

- 在做数据分析时，不能再仅仅依靠一小部分数据采样，而要利用所有的数据。
- 面对快速的、多源的、结构复杂的海量信息，我们一定要乐于接受，要不断扩大数据的分析量。
- 改变思考问题的方向，应关注事物之间的相关关系，而不再探求难以捉摸的因果关系。

随着科技的发展，我们可以处理的数据量已经大大地增加，而且未来会越来越多。甚至在某些方面，我们已经拥有了能够收集和处理更大规模数据的能力。

例如，ZestFinance 是一个利用“机器学习 + 大数据分析”为 payday loan 行业（发薪日贷款，类似高利贷的短期高利息借款）提供客户品质分析的平台。ZestFinance 平台与传统的分析方式不同，其可同时运营多个模型对所有的海量数据进行分析来判断各种可能性，再加上越来越多的数据来源和种类，然后这些信息被转化为几万个可对借贷者行为做出测量的指标，如诈骗几率、长期和短期内的信用风险和客户的偿还能力等。最后，各模型的结果被整合成最终结果，可在几秒内为用户提供最可靠的结果。

在数字化时代，数据处理变得更加容易、更加快速，人们能够在瞬间处理成千上万的数据。因此，面对过去小数据采样的思维方式，我们一定要及时转变过来，要利用所有的数据来思考问题。

5.1.2 充分利用这些数据

在大数据分析尚未被主流接受的时代，有超过三分之一的受访者表示，他们所在的企业结合大数据，实行了某种形式的先进的分析。在大多数情况下，他们仅仅采用非常简便的方法，例如数据抽样。

三百多年前，英国约克大学统计学家约翰·格朗特（John Graunt）采用样本分析法推算出鼠疫时期伦敦的人口数，这种方法就是后来的统计学。这个方法不需要一个人一个人地计算，可以利用少量有用的样本信息来获取人口的整体数据。

专家提醒

约翰•格朗特首次提出通过大量观察，可以发现新生儿性别比例具有稳定性以及不同死因的比例等人口规律，如男婴出生多于女婴；并且第一次编制了“生命表”，对死亡率与人口寿命作了分析，从而引起了普遍的关注。约翰•格朗特的研究清楚地表明了，统计学作为国家管理工具的重要作用，其他被认为是人口统计学的主要创始人之一。

在收集和分析数据都不容易时，随机采样就成为应对信息采集困难的办法。通过收集随机样本，人们可以用较少的花费做出高精准度的推断。因此，随机采样很快就被应用于公共部门和人口普查，甚至被用来在商业领域监管商品质量。随机采样取得了巨大的成功，成为了现代社会、现代测量领域的主心骨。

其实，随机采样一直都有较大的漏洞，它只是在不可收集和分析全部数据的情况下的无奈选择。统计学家们证明，采样分析的精确性随着采样随机性的增加而大幅提高，但与样本数量的增加关系不大。笔者认为这种观点是非常有见地的，为我们开辟了一条收集信息的新道路。

这就是我们要改变的思维，虽说随机采样是一条捷径，但它也只是一条捷径。随机采样方法并不适用于一切情况，因为这种调查结果缺乏延展性，即调查得出的数据不可以被重新分析以实现计划之外的目的。如果企业没有考虑逐步淘汰抽样调查和其他过去的所谓最佳实践的“神器”，他们真的是后知后觉了。

5.1.3 海量数据替代采样

在信息处理能力受限的时代，世界需要数据分析，却缺少用来分析所收集数据的工具，因此随机采样应运而生，采样技术（sampling technique）被誉为 20 世纪最伟大的成就之一。采样技术最通俗的解释是，从统计调查总体（population）中抽取样本（sample）进行调查，获取数据，然后对总体数量特征作出推断的技术，其流程如图 5-1 所示。采样的目的就是用最少的数据得到最多的信息。

如今，计算和制表不再像过去一样困难。感应器、手机导航、网站点击和 Twitter 被动地收集了大量数据，而计算机可以轻易地对这些数据进行处理。当我们可以获得海量数据的时候，采样技术也就随之失去了它的优势。

然而，采样一直有一个被我们广泛承认却又总有意避开的缺陷，现在这个缺陷越来越难以忽视了。采样忽视了细节考察。虽然我们别无选择，只能利用采样分析法来进行考察，但是在很多领域，从收集部分数据到收集尽可能多的数据的转变已经发生了。如果可能的话，我们要收集所有的数据，即将项目的整体数量当作样本来审核、测试、分析。这样，我们能对数据进行深度探索，而采样几乎无法达到这样的效果。

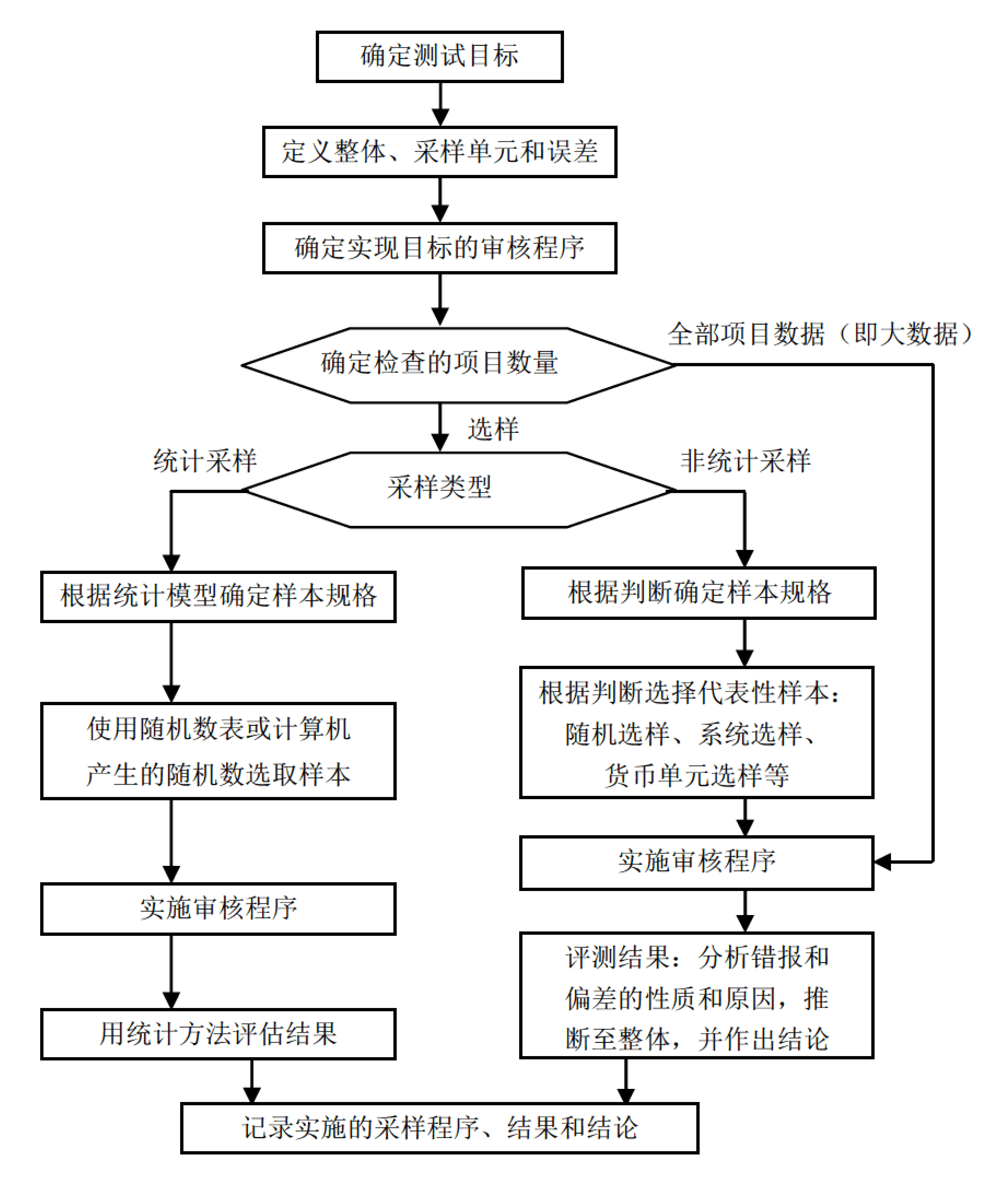


图 5-1　采样流程

通过使用所有的数据，我们可以发现如若不然则将会在大量数据中被淹没掉的信息。例如，信用卡诈骗是通过观察异常情况来识别的，只有掌握了所有的数据才能做到这一点。在这种情况下，异常值是最有用的信息，你可以把它与正常交易情况进行对比。这是一个大数据问题。而且因为交易是即时的，所以你的数据分析也应该是即时的。

随机采样只是一个暂时性的数据，随着你收集的数据越来越多，你的预测结果会越来越准确。数据处理技术已经发生了翻天覆地的改变，但我们的方法和思维却没有跟上这种改变。所以，我们现在要尽量放弃样本分析这条捷径，选择收集全面而完整的数据。

专家提醒

当然，想要用海量数据来代替采样也不是那么容易的，我们需要足够的数据处理和存储

能力，也需要最先进的分析技术。同时，简单廉价的数据收集方法也很重要。过去，这些问题中的任何一个都很棘手。在资源有限的时代，要解决这些问题需要付出很高的代价。但是现在，解决这些难题已经变得简单了。曾经只有大公司才能做到的事情，现在绝大部分的公司都可以做到了。

5.2 知己知彼，数据分析的演变与现状

以往的数据分析主要停留在结构化数据挖掘的阶段，例如移动、金融等企业内部的信息收集。目前，随着大量非结构化数据的产生，例如人的行为、富媒体、气候变化等内容，已经对业界提出崭新的挑战，一切事物都可以用大数据来分析。

5.2.1 大数据分析的商业驱动力

数据的应用与价值由来已久，随着互联网时代的发展，数据的开放为创新和价值生产的繁盛提供了一个平台，为商业不断打开了新的大门。新的商业模式、形态、传播该如何更好地利用数据，相信大家都仍在摸着石头过河。

因此，为了实现新的成本节省和增长计划，大量企业和机构在商业智能方案上投入重金，深入挖掘电子表格和各种不同系统（遗留系统、内部孤岛、客户关系、供应商、合作伙伴等）中的数据，以期获得接近实时的可操作分析结果（包括历史分析和未来预测）。

任何企业只要拥有正确的数据信息，就能较为精确地了解受众，知晓受众如何与你进行互动，知晓他们对于你的品牌有怎样的期待与回应。同时，数据还能帮助你更好地与受众进行针对性的互动与回应。因此，笔者认为，数据的关键价值在于其有效性。它能在定义受众市场、接触受众、与受众沟通等各阶段给予你有效的指引，并最终助推你的销售。

随着企业管理走向“信息驱动”，商业智能将成为企业信息计划的核心。大数据市场在未来五年将保持58%的惊人复合增长速度，会带来一场新的工业革命，如表5-1所示，而作为与大数据相关的商业智能平台和应用也将受益。

表5-1　大数据分析带来新的工业革命

进　程	第一次工业革命	第二次工业革命	第三次工业革命
时间	18世纪60年代～19世纪40年代	19世纪70年代～20世纪初	21世纪初
能源	蒸汽	电力	计算
材料	金属	化学	数据
工艺	机器制造	精密仪器	分析论证
特征	规模化	自动化	个性化

5.2.2 大数据分析环境的演变

“大数据”这个概念从 2008 年 9 月正式提出以来已经发展了 5 年多了，2012 年是大数据发展最快的时期，主要原因是，在 IBM 等多方厂商及政府的共同努力下，才使得“大数据”在中国变成一个流行概念。大数据分析的环境演变过程如表 5-2 所示。

表 5-2 大数据分析环境的演变过程

分析环境	大数据分析 1.0	大数据分析 2.0	大数据分析 3.0
数据来源	自身业务需求产生大量数据	收集与目标业务直接或间接关联的大量异质数据	对数据源的质量、价值、权益、隐私、安全等产生充分认识，出台量化与保障措施
分析论证	利用这些数据，通过深入的分析和论证，优化相关业务	建立复杂的分析和预测模型，产生针对目标业务的输出	数据运营商出现，数据市场形成，数据产品丰富，数据客（Dacker）活跃，促使分析论证方法进一步完善
数据价值	用数据指导决策	数据即决策	学术团体、企业和政府通过大量异质数据和数据产品产生科学、社会、经济等方面的新价值
应用案例	沃尔玛、亚马逊、百分点、豆瓣	Google Flu Trends（谷歌流感趋势）、ZestFinance、Google Powermeter（用电监测软件）	大数据实验室（BigDataLab）

1. 大数据分析 1.0——商业智能时代

在大数据分析 1.0 时代，数据管理已经有了实质性的发展，其能够客观分析和深入理解重要的商业现象，并且帮助管理者基于客观事实决策，而不是仅凭直觉。在商业实践中，生产流程、销售、客户交互乃至更多的数据，第一次被存录、整合和分析。

大数据分析 1.0 时代具有以下特点。

- 建立企业级数据仓库：最初，大公司凭借其雄厚资本可以定制数据系统；随后数据系统很快被商业化，可以由外部供应商以更通用的方式提供给更多公司。这就是企业级数据仓库的时代，系统可以捕捉数据，然后利用软件进行商业智能分析，最后可以进行数据查询和结果交付。
- 数据管理出现新问题：体量相对较小、流转速度较低时，数据组可以在数据仓库中分别存储并用于分析。但是，在数据仓库中进行数据准备和排序依然是一个难题。数据分析师往往要花大量的时间用在准备数据上，只剩下相对很少的时间用

在数据分析上。

- 数据分析的周期过长：数据分析师只能选择对几个非常关键的问题进行数据分析，因为分析需要数周甚至数月的时间，其过程艰难且缓慢。
- 大数据无法预测未来：作为商业智能最重要的部分——“数据汇报系统”只描述过去所发生的事情，既无法解释过去，也无法预测未来。

在大数据分析 1.0 时代，人们会把分析视为竞争优势的来源。但很少有人会使用类似“人才竞争”或“成本竞争”这样的方式来表述“分析竞争”。因此，企业应及时调整大数据分析的方向，将核心竞争优势放在更有效的运营基础上，也就是在关键节点上做出更好的决策，从而提高公司业绩。

2. 大数据分析 2.0——大数据时代

2005 年初，谷歌、eBay 等硅谷的互联网公司和社交网络开始大规模存储和分析新类型信息，尽管此时还没有产生“大数据”一词，但现实情况快速地改变了数据和分析师在企业内的角色。

大数据分析 2.0 时代具有以下特点。

- 数据量明显增大：大数据明显有别于系统内部产生的交易类“小”数据，它们是来自公司外部、互联网、传感器、各种公开发布的数据（例如人类基因组计划），还包括来源于音频和视频的数据。
- 出现新型商业模式：当大数据分析进入 2.0 时代，人们对于强大的新型分析工具的需求以及通过提供工具来获利的机会，很快就显而易见了。所有企业都忙于发展新能力和争取新客户。第一个“吃螃蟹”的企业很容易占得先机，获得令人印象深刻的宣传效果，并且会快速地研发出新产品。
- 创新技术如雨后春笋般涌现：例如，Hadoop 平台应运而生，其可以用来快速批处理大数据；新型数据库 NoSQL 可以处理相关的非结构化数据，使大量的信息可以在公有或者私有云计算环境里存储和分析；机器学习（半自动模型的研发、测试）则用于从实时动态的数据中迅速生成数据模型；色彩鲜明、立体效果的数据视觉化替代了单调的白纸黑字。
- 对分析人才提出了更高的要求：新一代的数据分析师被称为数据科学家，他们不仅要具备计算能力还要掌握分析能力。数据科学家已不再满足于被藏在公司内部，他们希望接触客户以开发新产品，并为公司出谋划策，甚至是创造新的商业形态。

3. 大数据分析 3.0——富化数据的产品时代

在大数据分析 2.0 时代，一些敏锐的观察者已经洞察到即将来临的下一个大时代——大数据分析 3.0 时代。

大数据分析 3.0 时代具有以下特点。

- 大企业纷纷介入大数据：例如，硅谷的大数据先驱公司开始投资面向客户产品、

服务和功能领域的数据分析。他们通过大数据分析吸引更多的访客登录他们的网站，这些办法包括更佳的搜索算法、朋友和同事推荐产品、购买建议以及针对性极高的定向广告等。

- 大数据的应用范围变得更广泛：如今，不仅仅是 IT 公司或者电子商务公司利用数据分析创造新产品和新服务，任何行业的任何公司都在这样做。无论企业属于制造类、运输类、零售类，还是服务提供类，这些商业活动都会产生大量的数据，任何设备、运输工具和客户都会留下痕迹，如果能够分析这些数据集，就可以更好地帮助积累客户和分析市场，帮助管理者做出适当的商业决策。
- 带来了全新的机遇和挑战：新的思维方式正在涌现，能掌握优势的新方法正在确立，新的参与者开始出现，竞争格局也随之发生变化，新的技术必须被熟练掌握，人才也应配置于最令人兴奋的新岗位上。那些能首先洞察到大数据分析 3.0 时代的公司，将会在引领行业变革的趋势中占据最佳位置。

5.2.3 大数据分析与处理方法

要知道，大数据已不再仅仅是数据量大，最重要的现实就是对大数据进行分析，只有通过分析才能获取更多智能的、深入的、有价值的信息。

如表 5-3 所示，是笔者对海量数据的处理方法进行了一个一般性的总结，当然这些方法并不能完全覆盖所有的问题，但是这样的一些方法也基本可以处理遇到的绝大多数问题。

表 5-3 大数据分析与处理方法总结

分析方法	适用范围	基本原理及要点
Bloom filter	可以用来实现数据字典，进行数据的重判，或者集合求交集	采用哈希函数的方法，将一个元素映射到一个 m 长度的阵列上的一个点，当这个点是 1 时，那么这个元素在集合内，反之则不在集合内
Hashing	用于快速查找、删除的数据结构，通常需要把全部数据放入内存	例如，在海量的日志数据中提取出某日访问百度次数最多的那个 IP，IP 的数目还是有限的，最多 2^32 个，所以可以考虑使用 hash 算法将 IP 直接存入内存，然后进行统计
bit-map	可进行数据的快速查找、判断、删除	使用 bit 数组（树状数组）来表示某些元素是否存在，即将原数据划分为多个区间，当要查询或更新某个数据或某段数据时，只需更新到各个区间不必细化到具体的各个元素
堆	可进行数据的快速排序	从海量数据中找出前 N（N 为比海量数据小的数）个数据，例如，从一亿个数据里，找出前 100 个最大的

续表

分析方法	适用范围	基本原理及要点
双层桶划分	用于确定数据的范围	面对一堆大量的数据我们无法处理时，可以将其分成一个个小的单元，然后根据一定的策略来处理这些小单元，从而达到目的。另外，如果需要用一个小范围的数据来构造一个大数据，也可以利用这种思想，相比之下不同的，只是其中的逆过程
数据库索引	大量数据的增加、删除、修改和查询	利用数据的设计实现方法，对海量数据进行增加、删除、修改和查询处理
Inverted index	搜索引擎，关键字查询	Inverted index（倒排索引）是一种索引方法，被用来存储在全文搜索下，某个单词在一个文档或者一组文档中的存储位置的映射。 以英文为例，下面是要被索引的文本： T0 = "it is what it is" T1 = "what is it" T2 = "it is a banana" 通过倒排索引方法就能得到下面的反向文件索引： "a": {2} "banana": {2} "is": {0, 1, 2} "it": {0, 1, 2} "what": {0, 1} 检索的条件"what"、"is"和"it"将对应集合的交集
外排序	大数据的排序	通常来说，外排序（External Sorting）处理的数据不能一次装入内存，只能放在读写较慢的外存储器（通常是硬盘）上。外排序通常采用的是一种“排序-归并”的策略。在排序阶段，先读入能放在内存中的数据，将其排序输出到一个临时文件，依此进行，将待排序数据组织为多个有序的临时文件。而后在归并阶段，将这些临时文件组合为一个大的有序文件，也即排序结果
tree 树	用于统计、排序和保存大量的字符串，经常被搜索引擎系统用于文本词频统计	利用字符串的公共前缀来减少查询时间，最大限度地减少无谓的字符串比较，查询效率比哈希表高

如今，越来越多的应用涉及大数据，这些大数据的属性，包括数量、速度、多样性等都呈现了大数据不断增长的复杂性，所以，大数据的分析方法在大数据领域就显得尤

为重要，可以说是最终信息是否有价值的决定性因素。

专家提醒

需要注意的是，尽管大数据已经有了长足的进步，但不要指望它能给予你长期的竞争优势。那些想要在新的数据经济中获得成功的企业，必须从根本上重新考虑如何利用数据分析为自己和客户创造价值。因此，我们要用全新的视角看待大数据“分析”的价值和作用，这意味着战略重点的转移。

5.3 企业管理中的大数据分析应用案例

关于数据分析对管理的重要性，在《孙子兵法》中已有深刻的描述：“夫未战而庙算胜者，得算多也。”意思是说，拉开战斗序幕之前，就已“庙算”（古时战前君主在宗庙里举行仪式，商讨作战计划和预测战争形势）周密，即充分估量了有利条件和不利条件，开战之后就往往会取得胜利。

同样，预测在企业中有重要的意义，在大数据时代，预测的准确度或许能够更上一个台阶，这将促进企业健康发展。因此，企业只有找到将数据科学与传统技能完美结合的方式，才能打败对手。不是所有的赢家都会将大数据用于其决策制定，但数据告诉我们，这样确实胜算最大。本节主要介绍大数据分析在企业管理中的应用案例，希望对读者有一定的启发和学习价值。

5.3.1 【案例】机场用大数据管理节省数百万美元

近日，美国里克哈斯本德阿马里洛国际机场（Rick Husband Amarillo International Airport）签署了 PASSUR 大数据解决方案合同，该方案旨在通过优化的机场管理为运营商提供最经济的运营。

PASSUR 公司研究机场的航班时间发现，大约 10%的航班实际到达时间与预计到达时间相差 10 分钟以上，30%的航班相差 5 分钟以上。为了提高服务质量，PASSUR 公司通过搜集天气、航班日程表等公开数据，结合自己独立收集的其他影响航班因素的非公开数据，综合预测航班到港时间。例如，由于天气原因造成延误时，应尽量让飞机在登机门处等候，而不是浪费燃油长时间在停机坪上等候。

里克哈斯本德阿马里洛国际机场航空部主管 Scott C. Carr 表示：“在当前的环境下我们非常注意的是，机场必须把两种价值作为最重要的事项：谨慎的财务监督和高效安全且经济的机场管理。PASSUR 是实现这些关键业务目标的理想合作伙伴。”

PASSUR 公司从美国联邦航空局处得到飞行计划、实时信息和每个航班的首个航点。随后工作人员会给每个航班分配 15 分钟进行排序。无论何种原因，如果空中交通

指挥塔台延长了计划时间，则所有的航空公司得到的配额时间都会相应减少。运营商可以在他们分配到的时间里更换自己的飞机。

目前，PASSUR 公司已经拥有超过 155 处无源雷达接收站，每 4.6 秒就收集一次探测到的每架飞机的一系列信息，这会持续地带来海量数据。使用 PASSUR 公司的服务后，里克哈斯本德阿马里洛国际机场大大缩短了飞机预计到达时间和实际抵达之间的时间差。航空公司依据 PASSUR 公司为他们提供的航班到达时间做计划，每年节省数百万美元。

专家提醒

企业管理学界因观点不同而分为众多派系，但是“不会量化就无法管理”的理念却是共识。这一共识足以解释近年来的数字大爆炸为何无比重要。有了大数据，管理者可以将一切量化，从而使公司业务尽在掌握中，进而提升决策质量和业绩表现。

【案例解析】：在进入大数据时代后，如何更好地利用信息爆炸时代产生的海量数据为管理服务和利用数据创造财富是不可回避的命题。成本领先战略、差异化战略、集中化战略是企业在市场竞争中可以选取的三大战略。在信息大爆炸时代，第四种竞争战略——大数据战略成为原三大竞争战略的支撑，其将改变企业决策、价值创造和价值实现的方式，如图 5-2 所示。管理决策日益基于数据和分析而作出，而并非基于经验和直觉，这对企业正确地制定发展计划与合理安排企业资源有重要的意义。

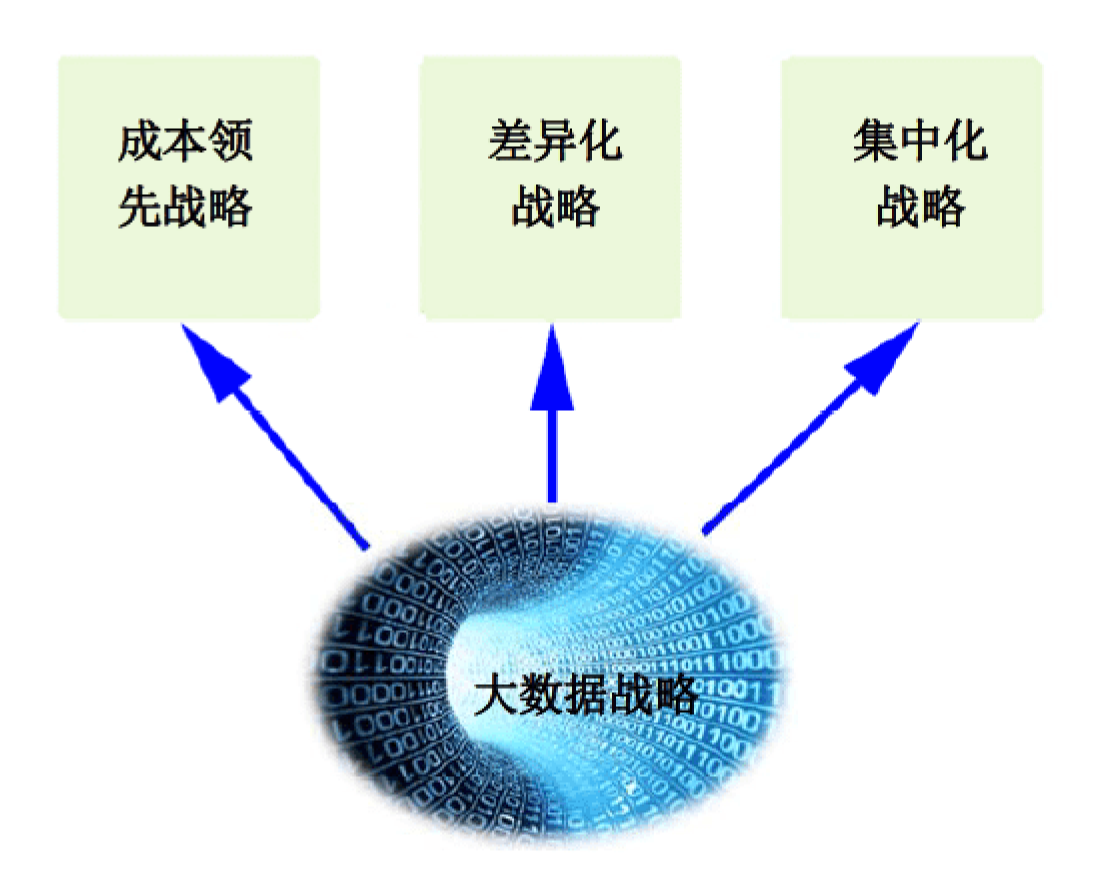


图 5-2　大数据战略支撑传统三大竞争战略

从上面的案例可以看出，对航空服务业来说，时间的精准就是优质的服务，尤其是航班抵达时间精准，这正好应了大数据战略的典型特点——预测变得更为精确。

5.3.2　【案例】国药集团打造全方位的管理模式

早在 2007 年，中国医药集团（以下简称“国药集团”）便启动了大数据商业智能的

建设，将集团的运营管理带入精细化管理的新时代。

国药集团在 10 余年的发展历程中，在原医药批发站的基础上一路并购，成就了今天拥有十大主营业务板块的规模最大的医药企业集团。与国药集团自身不断并购重组壮大的路径相似，集团的信息化建设也经历了不断演进的过程，最终形成了清晰的信息化战略。国药集团的信息化标准框架体系如图 5-3 所示。

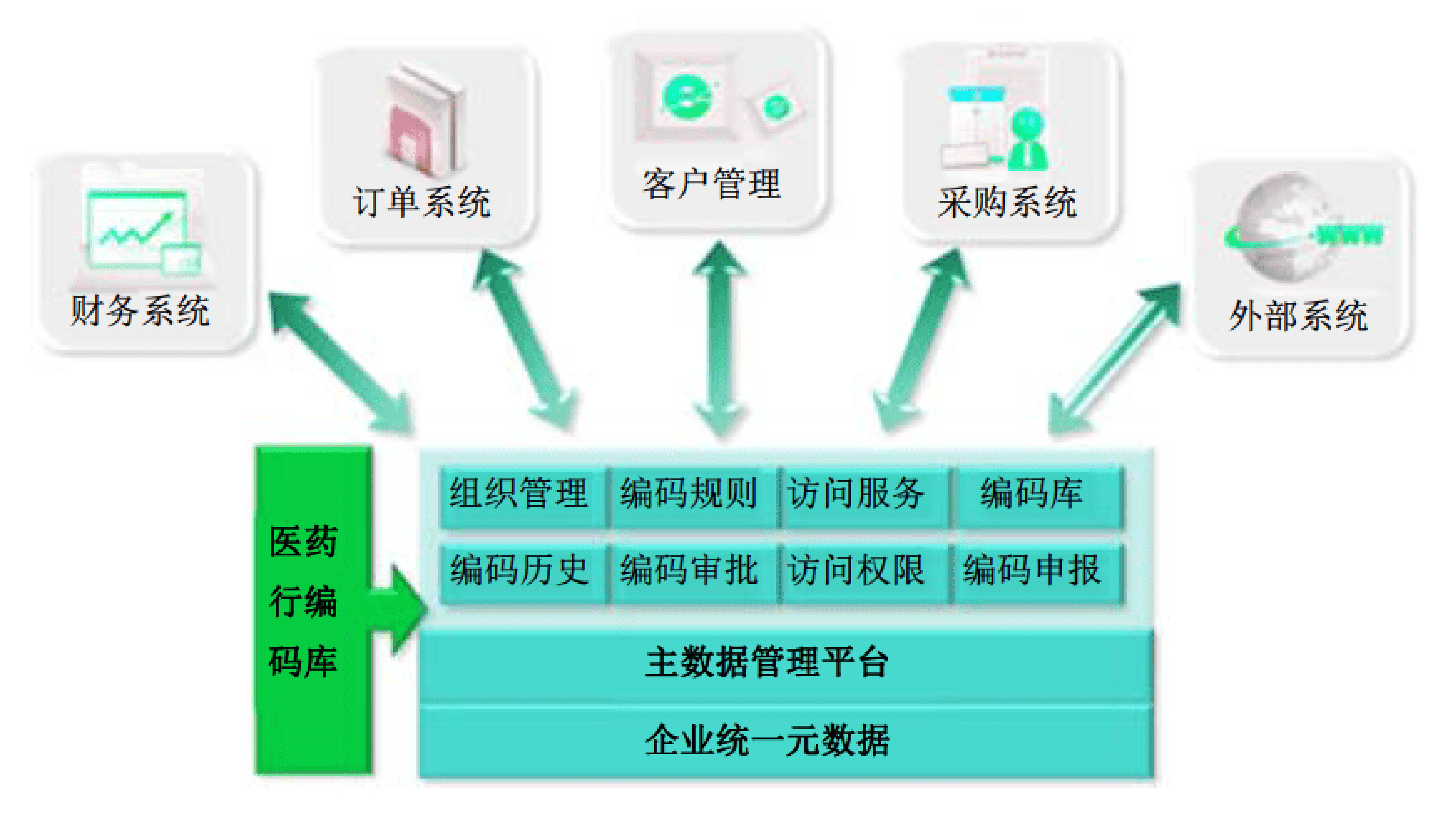


图 5-3 国药集团的信息化标准框架体系

国药集团正在全力推进集团 4 大平台——分销物流一体化营运平台、产学研一体化产业平台、国际化经营一体化平台、高效管控与融合协同一体化平台的全面建设，依托集团 4 大平台协同运作，促进集团 10 个核心业务——医药分销及物流配送、医药零售、生物制药、化学制药、现代中药、医疗器械、医药进出口及海外实业、化学试剂和诊断试剂、医药科研和设计、医药会展的全面发展，构成了一个完整的中央企业医药健康产业平台，实现了规模效益，推动业绩高速增长。

首先，国药集团要实现一体化运营，对下属二级和三级子公司进行全面的管控，及时了解子公司的运营、市场以及风险状况，就必须对子公司上百个系统的数据进行整合。只有将各异构的系统整合到统一的平台上才能形成集团式的管控和运营，这种整合必须通过 BI 去实现。

另外，国药集团还专门成立了运营管理部，借助 BI 系统对二级、三级企业的经营管理指标进行分析，从而优化运营管理。在 BI 系统建设的过程中，国药集团实现了企业管理的精细化，业务和数据的标准化。在业务升级过程中，将标准嵌入 ERP 实现循序渐进地融合。目前，国药集团还会将主数据管理系统、BI 和 ERP 集成到同一个信息平台，实现三个平台的互动，使 BI 产生的数据更加完整、准确、及时。

同时，国药集团 CIO 雷万云博士还指出：“云计算对于企业更多地将会发挥行动指南的作用。在信息化建设中，需要将云计算作为最终的目标，并且最大限度地挖掘 IT 的

价值，以真正引领业务的发展。”

这样的信息化思路，也为集团的信息化建设节约了大量资金。仅在 2007 年，在国资委信息化测评中，国药集团排名第 28 位，但投资额却在 100 位之后，单项测评中系统集成的方法论更是排名第一。

同时，主数据管理系统也作为进行业务规划的重要参考依据，减少业务规划过程中不必要的资源浪费，使业务结构更加优化。例如，主数据管理系统在 2008 年“5.12”抗震救灾医药物资调拨中发挥了重要作用，通过该系统工作人员能在第一时间内了解全集团医药储备库存情况，及时保障中央向灾区医药调拨，完成 3 亿共 800 吨的医药物资调拨任务，受到国务院、发改委、卫生部的表扬。

【案例解析】：通过主数据项目和 BI 项目，能够实时了解全面的各级业务部门经营状况以及管理统计分析数据，实现了集团对应收账款的及时管理和监控，有效防范集团风险，为国药集团管理层进行科学决策提供了重要参考依据，提高了决策效率，提升了决策质量。

笔者觉得，国药集团的精细化管理同时也是一种理念，更是一种文化。精细化管理是源于发达国家的一种企业管理理念，它是社会分工的精细化以及服务质量的精细化对现代管理的必然要求，是建立在常规管理的基础上，并将常规管理引向深入的基本思想和管理模式，是一种以最大限度地减少管理所占用的资源和降低管理成本为主要目标的管理方式。

现代管理学认为，科学化管理有 3 个层次：第一个层次是规范化，第二个层次是精细化，第三个层次是个性化。显然，大数据分析可以帮助企业完成精细化管理到个性化营销的过渡。

5.3.3 【案例】迪士尼乐园用大数据提升游客乐趣

迪士尼是孩子和童心未泯的成人的天堂，每个乐园里都有 100 多个项目，但每一个项目前等待的排队人群常常令人兴致大减。为此，迪士尼公司使用 10 多年的历史数据，结合天气、旅游等数据，预测每一条队伍每一天每一小时所需的排队时间，游客可以参考这个分析结果安排自己在园区内的游览次序。另外，迪士尼公司还收集了 Twitter 数据更新每一条队伍的排队等候时间，来处理突发的情况。

迪士尼公司的大数据策略，使每位游客平均每人节省 4 个小时，从而提升了游客们进园游玩的乐趣。

在大数据战略上取得初次成功后，迪士尼公园又准备投资数十亿美元打造度假计划系统 MyMagic，其核心支撑元素是它对每年到主题公园游玩的几千万旅客的数据进行收集的能力，这种技术是前所未有的。

MyMagic 系统将使迪士尼能够追踪游客去了乐园里的哪些地方、如何进行消费、在什么时候用餐和喜欢吃什么。迪士尼计划用这些信息制定出更细致和更个性化的营销方案，这样一来，该度假公园针对每位潜在用户所传达的信息和所制定的价格都是不同的。

MyMagic 系统的核心技术是腕带，官方命名为“MagicBands”（魔法带），其中嵌有无线射频识别芯片，其能与遍布迪士尼乐园的无线射频扫描设备进行通信，如图 5-4 所示。有些短距数据读取器安装在明显的位置，在购买纪念品或打开酒店房间时，游客可以在上面挥一挥自己的腕带。也有一些长距的读取器安装在隐蔽位置，游客无需进行任何操作，这些设备也能读取数据。

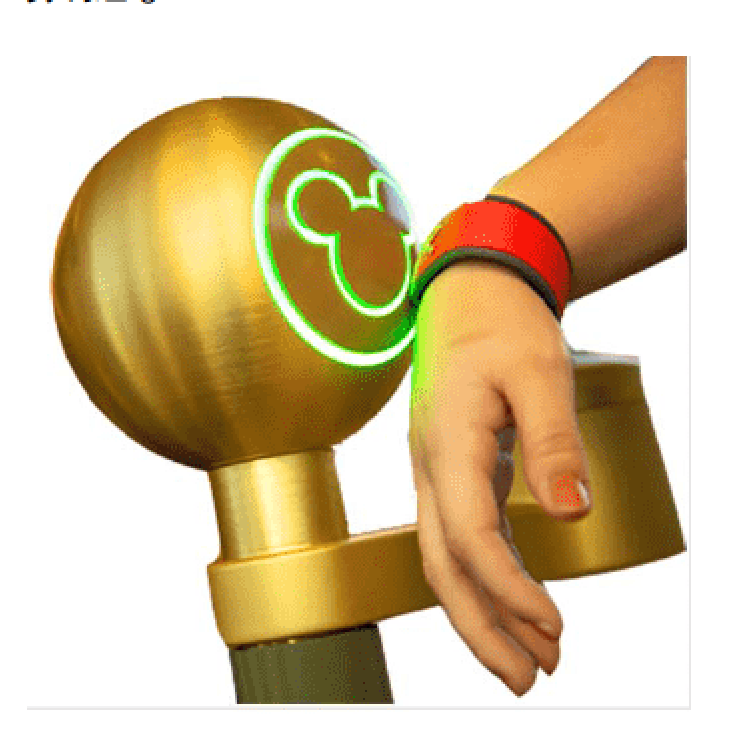


图 5-4　MyMagic 系统的核心技术——MagicBands

迪士尼将 MyMagic 的分析功能视为第二个增收工具。首要增收工具是鼓励游客提前安排好行程细节，以使他们在公园里呆更长时间以及通过更便捷的非现金支付手段来进行消费。例如，某个园区的一家餐厅在某个时间段有开店仪式，那迪士尼就可以通过 MyMagic 系统知道哪些在这个园区的游客在该时间段没有预订“FastPass”服务，然后向这些游客发送该餐厅的即时折扣。

【案例解析】：毋庸置疑，迪士尼是一个巨大的娱乐公司，但是当它涉及大数据平台，这位娱乐巨头看起来更像是一个初创公司。很多小公司，依靠坚强的意志和不凡的智慧，凭借一个小小的团队，使用 Hadoop、NoSQL 数据库和其他开源技术，完全能够创造出一个特有的大数据平台。

迪士尼能否有效地通过收集和利用数据来获利，很大程度上决定了该公司在 MyMagic 项目投入近 10 亿美元是否值得，以及它能否成为该公司的主题公园和度假区业务（年收入近 130 亿美元）的增长引擎。

从迪士尼的案例中可以看出，基于数据的竞争将提高组织的日常运营效率，找出可以省钱的地方和机会；基于数据的分析结果可提高决策速度和质量、增强预测能力，从而更好地理解客户和市场需要。因此，企业要学会计算数据的投资回报——数据价值和数据成本的比值。笔者可以毫不忌讳地说，降低数据成本和增加优质数据价值都是企业管理者要关心的方向。

5.3.4 【案例】Farmeron 用大数据促成农业增产

农业市场的潜力是巨大的，据国外调查统计可知，全球范围内中型企业规模农场的市场价值已经达到 120 亿美元，但截止至今，这些农场仍大多依照的是传统陈旧的运行系统。

Farmeron 是美国加州山景城的一家创业公司，Farmeron 看到了传统农业生产管理中的诸多不足，试图颠覆传统，成为世界上首批农业 SaaS（Software-as-a-service，基于互联网提供软件服务的软件应用模式）公司之一。Farmeron 开发了一款类似于 Google Analytics 的数据跟踪和分析服务产品，旨在帮助全世界农民在线管理其产品信息，使用统计方法进行自动农场运作状况分析，帮助农民提高工作效率。

Farmeron 打造了一个分析工具包，农民可在其网站上利用这套工具，记录和跟踪自己饲养的动物的情况（饲料库存、消耗和花费，每头动物的出生、死亡、产奶等信息，还有农场的收支信息）。就像我们在 Facebook 或者 Twitter 上有一个主页一样，每个动物也都有一个自己的页面，这可以让农场主不仅看到整个农场的表现，还可以看到每一只动物的情况，如图 5-5 所示。



图 5-5　农场管理工具 Web 页面

Farmeron 的创始人马提亚・可匹克（Matija Kopić）来自克罗地亚一个农场主家庭，不过他最终与父母走上不同的道路，成为一名程序员，他希望用一种现代化的方式来减轻农场主的工作负担。多数软件创业公司的创始人整日对着电脑测试代码，马提亚・可匹克却常在畜棚度日。

马提亚・可匹克专注于使分析报告和操作界面便捷易用，像个人理财网站 Mint 一样省心。过去一位奶牛场经理需要花几天时间来输入和分析几个月来的奶牛进食与医疗数据，如今结论立等可取。

少年时期的马提亚·可匹克在制作奶酪上很有一套，但这一兴趣并未影响他的另一项激情——到萨格勒布大学攻读计算机科学。如今，他带着自己的创业公司 Farmeron 回到了这片土地。

自 2011 年 11 月成立至今，Farmeron 已在 14 个国家建立起农业管理平台，目前已有超过 600 家企业化农场使用该产品，其中 45%都位于北美，最大的一家拥有 4000 头牲畜。2013 年 5 月，Farmeron 又与在 30 多个国家开展业务的大型德国设备商 Neelsen Agrar 达成协议，由后者向客户销售 Farmeron 软件。另外，Farmeron 已经在其发起的种子轮融资中获得了 140 万美元的投资资金。

一位管理着一个拥有近 400 只牛的奶牛场兽医表示，Farmeron 帮助他满足了动物信息追踪和销售方面的需求，该工具还有助于及时向保险公司汇报牲畜死亡情况。兽医还用 Farmeron 管理日常饲料配给及饲料采购，并不断进行微调，这相当重要，因为饲料成本占到了他这个奶牛场总成本的 70%。“只要能省一点钱，我们都努力去省，”兽医表示，“我经常能够看到饲料中某个成分不符合计划，从而可以迅速作出反应。”

【案例解析】：农民们一向拥有海量信息，但他们既没有可用于分析的工具，也没有接受过相关训练。在本案例中，由于 Farmeron 从很多农场那里收集数据，它可以就何种方法有效得出适用范围很广的结论，并建议如何提高产量。Farmeron 帮农民把支离破碎的农业生产记录整理到一起，用先进的分析工具和报告，帮农民达成农业生产计划。目前，世界人口总数已突破 70 亿，这也就迫使农业必须变得更加高效，而这也正好能够促进 Farmeron 的发展。

使用大数据分析，还可以帮助农场针对市场上竞争对手的市场策略进行实时的反应并调整价格。笔者认为，Farmeron 可以使用大数据来为农场提供个性化的在线服务，满足个性化的需求，这样销售额和利润的增长会更加见效。

5.3.5 【案例】西尔斯着眼于大数据以降低成本

全球 500 强企业之一的西尔斯控股公司（Sears Holding），这家几乎与西方现代零售业同龄的老古董公司，曾经雄居美国零售业榜首近一个世纪。但是，最近几年，这个零售巨头的日子却是江河日下，前途不容乐观。

有两方面的原因导致西尔斯的规模下滑：一是西尔斯公司近几年一直在大规模地关店，但同时也有新店开张，而且整体门店数量有波幅的上涨；第二点，也是让西尔斯更绝望的，就是其门店可比销售负增长，而且近几年全部出现负增长。

为了改变企业管理方式，抑制不良形势的继续发展，西尔斯控股公司首席信息官 Keith Sherwell 近期为该零售企业规划了一幅全面的技术革新蓝图，而这幅蓝图要想成为现实，则要依赖于 Hadoop、开源以及进一步削减管理维护成本。

西尔斯公司收集其专售的三个品牌——Sears、Craftsman、Lands'End 的客户、产

品以及销售数据，从这些海量信息中挖掘价值。大数据潜在价值巨大，但挖掘和分析这些数据的困难也很大。

- 数据量庞大：首先需要对这些数据进行超大规模分析，且这些数据分散在不同品牌的数据库与数据仓库中，不仅数量庞大而且支离破碎。
- 分析时间长：西尔斯公司需要 8 周时间才能制定出个性化的销售方案，但往往做出来的时候，它已不再是最佳方案了。

西尔斯公司首席信息官 Keith Sherwell 近期作了一份关于大型零售集团企业的技术革新计划。Sherwell 的规划思路来自于一次由 Cowen 公司主办的关于大数据的公开会议。显然，Cowen 公司的分析师 Peter Goldmacher 将大数据的发展规划草图有效地传达给了 Sherwell，并被带进了西尔斯公司。

此后，西尔斯公司开始使用集群（cluster）收集来自不同品牌的数据，并在集群上直接分析数据，而不是像以前那样先存入数据仓库。为了避免浪费时间，西尔斯公司先把来自各处的数据分析之后再做合并，这种调整让公司的推销方案变得更快、更精准。

专家提醒

简单地说，集群（cluster）就是一组计算机，它们作为一个整体向用户提供一组网络资源。其中，单个的计算机系统就是集群的节点（node）。

【案例解析】：最好的大数据供应商，是那些能将数据以最合适的形式呈现出来的供应商。从本案例可见，西尔斯公司力求拥有零售行业中规模最大的 Hadoop 集群，该企业在开源上下了很大的赌注。

传统的企业管理流程是出现问题、逻辑分析、找出因果关系、提出解决方案，从而使问题企业成为优秀企业，这是逆向思维模式。大数据竞争战略咨询流程是收集数据、量化分析、找出相互关系、提出优化方案，从而使企业从优秀到卓越，这是正向思维模式，如图 5-6 所示。

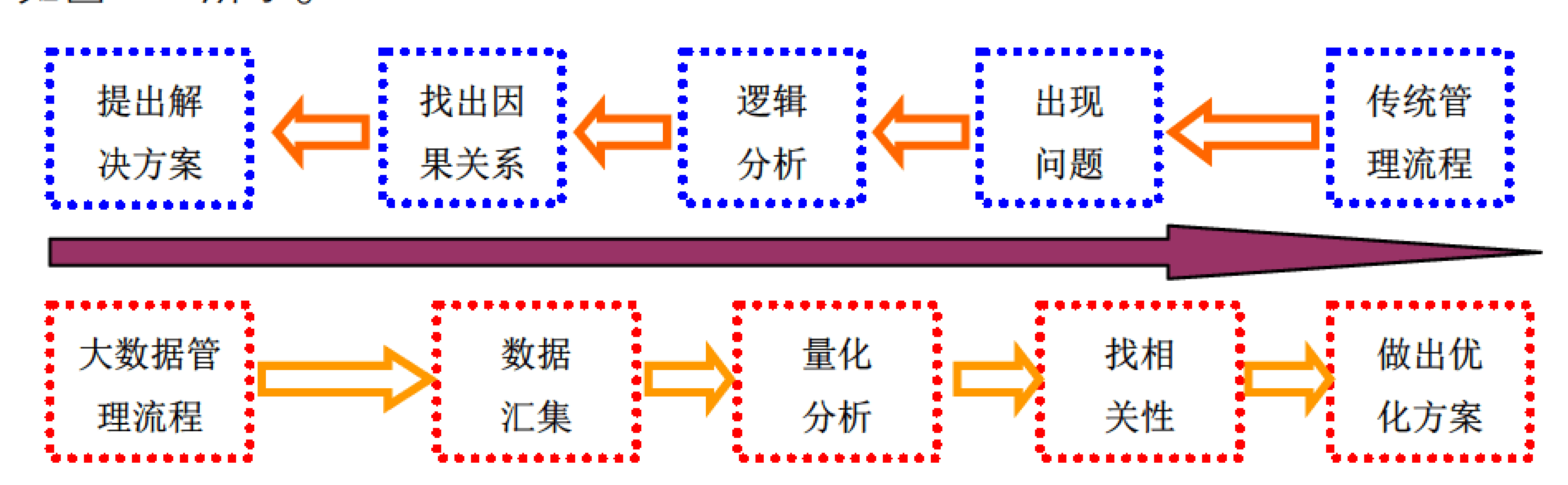


图 5-6　大数据管理与传统管理的思维模式区别

笔者认为西尔斯公司不是星星点点的个案，而是代表了整个商业的一次根本性经济转型。笔者确信，大数据运用带来的这一转型已经触及了商业活动的方方面面，没有谁能置身其外。

5.4 能源管理中的大数据分析应用案例

众所周知，自从三次科技革命以来，能源成为了国家经济的命脉。然而，地球上的能源是有限的，于是在各个大国之间引发了一些与石油有关或纯粹是为了石油的战争。为了争夺对世界资源与能源的控制权，导致了两场世界大战的爆发。

- 第一次世界大战中，31 个国家 15 亿人口卷入了战争，伤亡人数达 3100 万，其中死亡 1000 万人，军费支出与战争损失共计 3877 亿美元。
- 第二次世界大战中，7 年的战争中有 60 个国家参与，总伤亡人数达 9000 万人，死亡了 5000 万人，直接军费支出 1117 亿美元，物质损失 3 万亿美元。第二次世界大战后美国和苏联两个超级大国为了争夺资源与能源展开了 40 多年的冷战。

如今，对中东石油、南非的黄金和金刚石、扎伊尔的铜矿等资源的争夺战还在延续，可以说，能源战争将愈演愈烈。能源费用与日俱增，这促使很多商业机构和行业企业开始考虑通过技术节省能源开支。要想准确预测能源消耗并采取及时有效的节能措施，需要进行大量的数据分析。本节主要介绍大数据分析在能源管理中的应用案例，希望对读者有一定的启发和学习价值。

5.4.1 【案例】用“大数据”预测风电和太阳能

近日，IBM 宣布了一项先进的结合了大数据分析和天气建模技术的能源电力行业先进解决方案，将其命名为“混合可再生能源预测”（HyRef），旨在帮助全世界电力能源行业，提高可再生能源的可靠性。

HyRef 技术采用了天气建模能力、高级云成像技术和云图拍摄机来追踪云层运动，同时使用安装在涡轮上的传感器对风速、温度和风向进行监测，如图 5-7 所示。通过与分析技术相结合，这个以数据同化（Data-Assimilation）为基础的解决方案，能够为风电厂提供未来一个月区域内的精准天气预测或未来十五分钟的风力增量。

另外，HyRef 可以通过整合这些当地的天气预报情况，预测每个单独的风力涡轮机的性能，进而估算可产生的发电量。HyRef 充分利用大数据的洞察力，使能源电力公司可更好地管理风能和太阳能的多变特性，更准确地预测发电量，并且使其可以被复位导向到电网或储存。同时，HyRef 也可使能源组织更好地同时使用可再生能源与其他传统能源，例如煤炭和天然气。

HyRef 由 Deep Thunder 等创新技术发展而来，是气候建模技术领域内的一项高新成果。由 IBM 所开发的 Deep Thunder 技术可为特定区域内的气候状况提供高清微型预测，覆盖范围可从单一城市扩大至整个省份，并可达到平方公里的计算精确度。Deep Thunder 与商业数据结合后，可为商业用户和政府提供定制化服务，更改路线并加装设

备，来降低重大天气事件所带来的影响，从而降低成本、提高服务质量，甚至是避免人身危险，将重大气象引发的意外事件几率降到最小。

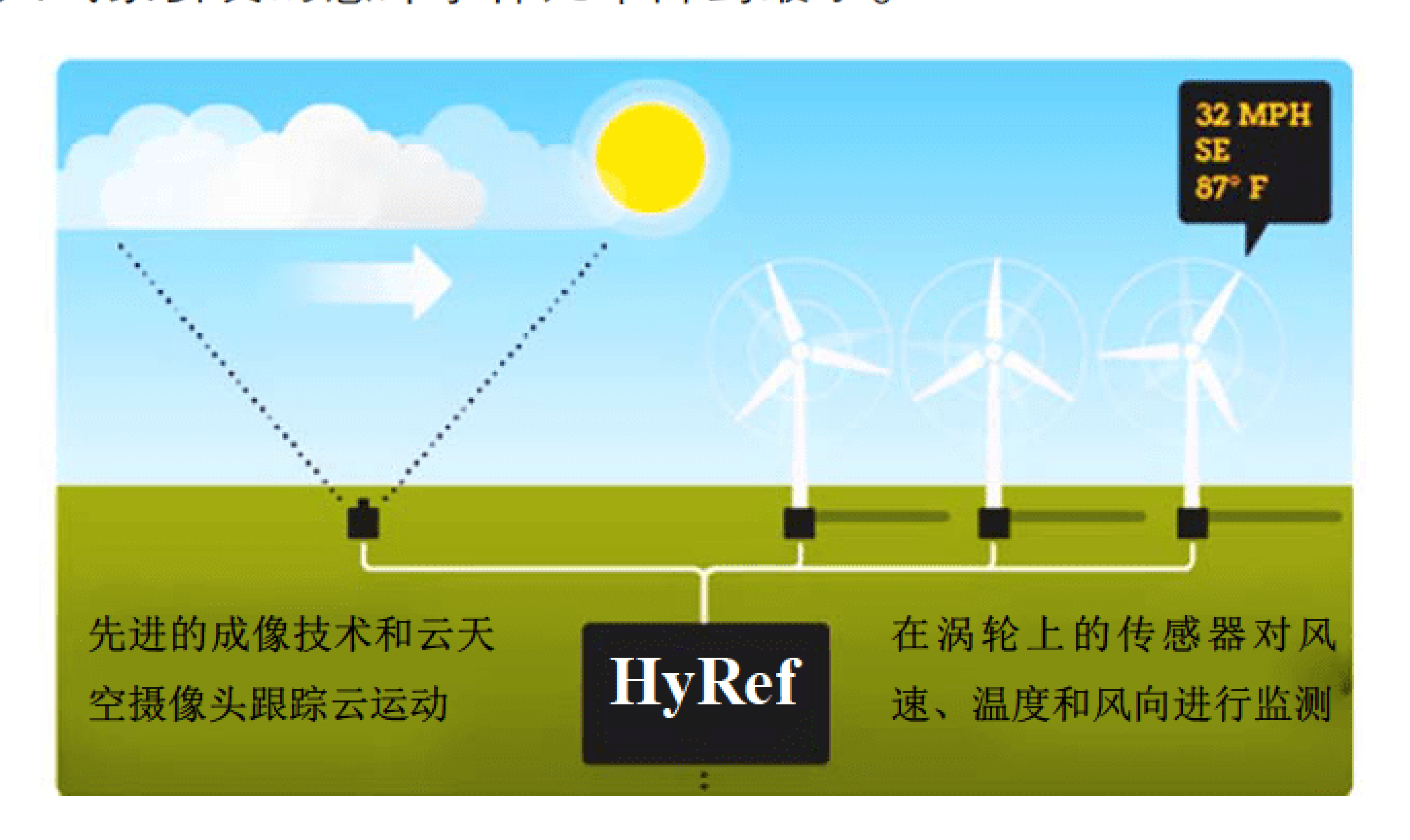


图 5-7　HyRef 的基本原理

【案例解析】：在本案例中，使用分析结果并有效利用大数据，将可使电力公司有能力应对可再生能源的间断特性并对太阳能和风能的产量做出合理预测，这是一种前所未有的创新模式。HyRef 使能源电力公司可将更多的可再生能源并入电网，减少碳排放量，给消费者与企业提供更多的清洁能源。

目前，全球的能源公司都在使用一系列策略将可再生能源集成到各自的系统中，以期在 2025 年前达到可再生能源在整体能源投资组合中 25%的占有率的基本目标。笔者可以预见，不久的将来，随着工业化和信息化的融合，大数据将深刻地影响能源行业和能源企业。

专 家 提 醒

据悉，中国国家电网（SGCC）所属的国家冀北电力有限公司（SG-JBEPC），正在使用 HyRef 整合可再生能源并将其并入所属电网中，该项目有助于实现中国“减少对化石燃料依赖”的 5 年计划目标。可见，我国的能源企业也在逐步实现数据管理化。

5.4.2　【案例】电力增长情况反映宏观经济形势

当今社会，电气化水平的提高使得各种经济活动几乎都离不开电，电力是国民经济发展中重要的生产资料及人民生活中必不可少的生活资料，与经济密切相关。用电量的变化及电力消费结构的变化也反映出经济运行及结构的变化。电力行业专家表示，全国用电量与 GDP 关联极其密切，大约是 1:1.2 的关系，用电量的持续下滑意味着经济增长乏力，用电量持续上涨意味着经济处于上行通道、发展势头良好。

因此，在官方 GDP 统计数据公布之前，人们寄希望于通过全社会用电量的变化来观察经济走势。针对这一要求，华东电网有限公司（以下简称“华东电网”）投运了一

套应用 BI 的调度生产统计分析系统，应用该系统可以看到累计到当天的年用电量情况，同比、环比也一目了然。这也就意味着，人们可以看到累计到前一天的 GDP 同比、环比情况。如表 5-4 所示，为华东电网近年来的大数据部署方案。

表 5-4　华东电网近年来的大数据部署方案

时　　间	大数据部署方案	具 体 内 容
2007	搭建电力市场分析系统	华东电网首先在电网调度中心部门、交易中心部门进行部门级试探性应用。调度部门是一个数据密集型的部门，如何有效利用这笔宝藏是门艺术
2008	建立数据集成平台	伴随着国家电网 SG186 工程的全面实施，以及智能电网的探索性工作开展，华东电网开始深化数据集成标准的研究，形成企业内部的信息集成标准，并在此基础上建立集成平台，完成了数据集成的工作
2009	建立企业级数据仓库	华东电网建立了全公司的企业级数据仓库，加大了集成的数据信息范围，并在这个企业级数据仓库的基础上，完成了一部分面向企业管理层的 BI 应用开发
目前	横向拓展大数据应用	在华东电网的各个信息系统中，传统报表和应用 BI 的数据分析是并存的，前者的普遍应用在生产运营中发挥了很好的作用，后者则提供了新的视角和实现手段。在未来的一段时间内，华东电网将会把 BI 应用进一步扩展到企业的其他业务部门

【案例解析】：通过电力增长情况来发现经济运行的一些情况，及时预测宏观经济形势，这应该是在所有经济数据中，最具实时性的指标之一了。

其实，对企业管理层而言，大数据的应用主要体现在战略导向上。在本案例中，华东电网将大数据与其他可视化手段进行了有机组合，以企业战略层最为关注的电网规划、电网建设、同业对标等指标为导向，打造了面向经营管理层的“全景可视化辅助决策分析系统”。另外，统一信息平台和全景可视化辅助决策分析系统是同步展开的，前者从下往上，后者从上往下，共同满足不同用户的数据需求。

专 家 提 醒

需要注意的是，用电量数据并不能简单代替 GDP，因为用电量数据的波动性强于 GDP，在经济增长处于高位的时期，用电量增速大于 GDP 增速；当经济处于下行周期时，用电量增速下降的幅度会大于 GDP 增速下降的幅度。

5.4.3　【案例】石油公司用大数据追求最大利益

美国阿美拉达赫斯公司（Amerada Hess Corp，Hess）是一家综合石油公司，总部

设在美国纽约，主要从事勘探、生产、购买及销售原油和天然气，勘探和生产活动遍布美国、英国、挪威、丹麦、印度尼西亚、泰国及其他国家。

Hess 公司的 CIO Gary Lensing 表示："我们做的任何事都是数据说了算；价值的量化亦全仰赖资料。"在过去几年里，Hess 不断地致力于建立基于大数据分析平台的 BI 系统，以尽可能地实时追踪从勘探到生产这条价值链上的所有数据。

Hess 公司的 BI 系统旨在能够查看 Hess 在挪威、丹麦、英国、美国、泰国以及非洲各地所有资产的活动。例如，非洲赤道几内亚的 4 座油田产量，今天是否达到预期？美国新泽西州的炼油厂，是否已用最大产量在生产？或是能否在月底前出产更多桶石油？某个时间点内，其 1370 家的加油站销售情况如何？

财务分析方面，Hess 主要是运用 Hyperion 的工具进行分析。为了估计他们的油井可以出产多少石油或天然气，Hess 在该地区为油田地形开发了一套模型，如图 5-8 所示。为了查看油井生产的特征，Hess 运用了在制药公司很普遍使用的一款工具——Tibco 公司的 Spotfire 产品，让分析人员可以通过图形、图表，以及其他图像来显示数据，用户于其中查询即可深入分析这些数据。



图 5-8　油田地形模型

Hess 还安装了 OSIsoft 绩效管理软件，用于收集操作上面的资料，例如，用来衡量钻井平台与储油槽的运作效率如何。同时，Hess 每天都通过 FTP 传输、接收其合资企业上载的绩效报表。

如今，钻井平台的工作人员能够与公司总部的人员进行实时对话，并且处理同一笔数据。例如，一位在美国德州休斯敦的工程师，可以对位于西非地区的钻井活动进行监控，查看钻头钻入海床时有无任何异常，并且可以通过卫星传输数据给休斯敦的工程师，他们可以检视此可视化的数据，然后发送电子邮件提出如何调节该机器的措施。

【案例解析】：理论上来说，大数据平台为 Hess 带来了更大量与更快速的产出，这意味着 Hess 可以在市场价格高涨时，更快速地卖出更多的原油或提炼产品。

"石油工业是信息工业"，很少有其他工业领域像石油工业这样更依赖于数据。对油气资源的认识和掌握主要通过大量的数据来实现，"大数据"往往意味着"大油气"，

通过对数据的挖掘和应用，可以提高决策的准确性和全面性，实现新的油气增产。就像石油、矿山对于工业革命一样，大数据正在成为信息社会最重要的战略资产，散发出令人难以抗拒的财富气息。

5.4.4 【案例】大数据管理更准确、一致、及时

农业部信息中心是国家农业部直属事业单位，负责承办农业部网站，是农业部的信息集散中枢和网络中枢，其主要任务是为农业部和党中央、国务院进行农业决策与管理提供信息服务，为各部行政机关提供通信、网络和信息支持，为全国农业系统及其农产品生产者、经营者提供信息社会化服务。

为了更好地利用数据资源，农业部信息中心决定建立统一的数据仓库平台，将各业务系统数据进行面向分析的整合，为管理人员提供更准确、一致、及时的决策支持信息。经过细致的考察、调研和选型，农业部信息中心选择了 CA 公司的数据仓库解决方案。

专家提醒

CA 公司（CA Technologies，CA）是全球最大的 IT 管理软件公司之一，其专注于为企业整合和简化 IT 管理。CA 公司创建于 1976 年，总部位于美国纽约长岛，服务于全球 140 多个国家的客户。

CA 公司是全球领先的 IT 管理软件和解决方案供应商，其产品和技术涵盖 IT 的所有方面，从主机到分布式系统，从虚拟化到云。农业部信息中心的 CA 数据仓库项目可以分为数据仓库的设计、构造和前端展现 3 个阶段，其中，每一个阶段都采用了不同的工具，如表 5-5 所示。

表 5-5　农业部信息中心的 CA 数据仓库项目流程

流　程	项 目 内 容	主 要 功 能
第一阶段	设计数据仓库	用户需求分析及数据仓库模型设计
第二阶段	构造数据仓库	采用 CA 的数据转换工具 Advantage Data Transformer，支持各种关系数据库和 ODBC 数据源，对数据进行完整的抽取、映射、转换，提供完善的编程能力以定制复杂的转移规则
第三阶段	数据仓库前端展现	CleverPath OLAD 在线分析、报表和决策支持系统

农业部信息中心数据仓库项目包括以下软件工具和模块：数据仓库建模工具、数据仓库数据转移工具、数据仓库 OLAD 分析及前端展现工具、决策支持/高级领导信息系统构造工具、生产报表工具。

目前，农业部信息中心数据仓库项目已验收成功并正式投入运营。CA 公司数据仓

库解决方案对农业部的业务管理作用是显而易见的，农业部信息中心已经充分利用数据仓库，建立起农产品贸易数据集市、农产品价格数据集市和气象数据集市，同时，定期由数据仓库自动生成农产品贸易信息和价格信息，在互联网上发布，为广大的中国农业信息网用户提供便利的信息服务。

【案例解析】：在本案例中，农业部信息中心通过数据仓库系统，可以使各级管理人员、信息分析人员非常方便地采用 C/S 和 B/S 模式对数据进行分析和查询，其快速的分析过程、准确可靠的分析结果，使工作人员的工作效率和质量大为提高。

专家提醒

C/S（Client/Server）模式是 20 世纪 90 年代管理信息系统（MIS）中较为先进的技术，C/S 应用系统基本运行关系体现为“请求/响应”的应答模式。每当用户需要访问服务器时就由客户机发出“请求”，服务器接受“请求”，并“响应”，然后执行相应的服务，把执行结果送回给客户机，由它进一步处理后再提交给用户。

随着信息技术的发展，C/S 模式已无法完全满足人们的需要，而且静态网页也无法提供充分的交互功能，动态信息发布相对较困难，这就需要将数据库与 Web 服务器连接起来，供用户查询或更新，而发布动态信息还可以简单到只需改动一下数据库的若干记录或字段就可以实现。这样，B/S（Browser/Server）模式在管理信息系统中开始大量应用。B/S 结构体系多了 Web 服务器，用户使用 Web 浏览器访问 Web 页，从数据库获取的信息能以文本、图像、表格或多媒体对象的形式在 Web 页上展现，用户通过 Web 页上显示的表格与数据库即可及时进行交互操作。

5.4.5 【案例】大数据帮助消费者提高能源效率

Pecan Street 是一个非营利性组织，由德克萨斯大学、相关技术公司和公用事业提供商组成，它们共同协作在智能电网技术领域进行测试、试运行和商业化运营工作。Pecan Street 的核心工作是研究一种终端设备到云的架构，其能够捕获多个来源的数据，并进行存储以供分析和可视化之用。

Pecan Street 主要通过一些系统收集电力数据，还通过使用无线网关的公用事业量表收集燃气和水的数据。例如，Pecan Street 通过记录消费者的行为，会自动修改其家庭中的环境控制方式（如空调系统等），或者调整其能源信息查看方式等。Pecan Street 还计划收集来自高级恒温器、家庭自动化系统、家庭安保系统、运动探测器以及新能源技术（如太阳能板和电动汽车充电站）的数据。

Pecan Street 采用了 EMC 公司的 Greenplum 大数据解决方案。Greenplum 系统采用了大量并行处理（MPP）架构，可帮助 Pecan Street 利用针对结构化和非结构化数据的模块化解决方案来处理和分析数据。

Pecan Street 除了要寻求合适的大数据分析方法外，收集的数据完整性也是一大问题。例如，数据系统中的无效信道或者居民宽带连接中断都会提供不可靠的值。因此，Pecan Street 通过生成已知完好数据的合格数据集来解决此问题，将这些数据标记为极高质量，并指引研究人员使用这些数据。

当前，Pecan Street 已经通过德克萨斯州奥斯丁市 Mueller 社区 200 多个家庭中的传感器系统，收集了近两年的能耗数据。利用大数据分析，Pecan Street 可以更好地了解人们的能源消费方式及其希望的能源管理方式。此外，Pecan Street 可以向公用事业公司提供洞察业务，帮助他们在电网改造领域进行最佳投资。

【案例解析】：大数据分析可产生大量的价值，正如大多数有价值的工作一样，大数据项目在一开始可能会困难重重，但它绝对值得我们投入时间和精力去挖掘其中的价值。在本案例中，Pecan Street 项目的主要目的是推动在消费者能源管理领域发现新的产品、服务和经济机会。

笔者认为，Pecan Street 的研究将可以向人们提供管理和减少其能耗的知识和工具，以帮助消费者提高能源效率，使其家庭生活更舒适。此外，公用事业公司将能够利用此类数据更好地管理电网，并投资更佳的基础设施改造工作。同时，笔者建议 Pecan Street 以及其他相关企业再接再厉，利用大数据分析进一步完善其“智能电网”系统，解决电网运营的 4 大核心问题，如图 5-9 所示。



图 5-9　电网运营的 4 大核心问题

6 案例：摆脱大数据风险

学前提示

我们在谈论大数据的美好前景时，当然不能完全忽略它可能带来的风险。很多人目前只关注大数据化带来的后果，如信息安全，而没有关注如何看待大数据本身的风险。本章将就当前尤其国内技术环境下，进入大数据时代所面临的风险和存在的问题做简要分析。

要点展示

- 问题凸显，大数据存在 5 大风险
- 步步小心，大数据项目 7 大误区
- 踏雪无痕，彻底逃离大数据监视
- 有备无患，做好大数据风险管理
- 大数据风险管理应用案例

6.1 问题凸显，大数据存在 5 大风险

对于大多数企业来说，大数据已经成为左右战局的决定性力量，安全风险也随之更加凸显。企业已经搜集并存储了所有的数据，接下来他们该干些什么？他们如何对这些数据进行保护？而且最为重要的是，他们如何安全合法地利用这些数据？

当然，任何事物都是把双刃剑，大数据正在变成生活的第三只眼，它敏锐地洞察却也正监控着我们的生活。想一想，亚马逊监视着我们的购物习惯，百度监视着我们的网页浏览习惯，微博似乎对我们和我们朋友的关系无所不知。

大数据的确改变了我们的思维，更多的商业和社会决策能够“以数据说话”。不过除了这所有利好，如何让大数据不侵入我们的隐私世界，也是与之伴生并需严肃考虑的问题。

6.1.1 风险 1：个人隐私泄露

正被美国全球通缉的斯诺登不久前“闯入”上海一场大数据研讨会。确切地说，研讨会的多位发言者都提到被斯诺登捅破的“棱镜门”。从纯技术角度观察，“棱镜”是一个典型的通过分析海量通信数据获取安全情报的大数据案例，但它也引发了思考：大数据时代，个人隐私该何处安放？

在大数据时代的背景下，你可以想象一些场景，如图 6-1 所示。



图 6-1　大数据时代背景下的隐私泄露途径

在大数据的时代背景下，一切都数据化了，我们平常上网浏览的数据，我们的医疗、交通、购物数据，统统都被记录下来，这就是大数据的起源。在这个时候，我们每个人

都成了一个数据产生者，数据贡献者。大数据的神奇之处在于，通过对大数据的分析，其他人甚至能够在很大程度上精确地知道你是谁。

人的行为看似随机无序，但实际上是存在某种规律的。社交网络如此发达的今天，大数据把人的行为进行放大分析，从而能够相对准确地预测人的性格和行程。所以，不排除有这样一种可能：在忙完了一天的工作之后，你还没有决定要去哪儿，数据中心却先于你预测了你接下来的目的地。

例如，在央视“3・15”晚会上，安卓手机软件窃取用户隐私信息的情况得以披露。然而，这仅是冰山一角。2013 年前 3 个月，金山手机毒霸检测到恶意侵犯用户隐私的安卓软件共计 2.3 万个，每天有 41 万部安卓手机能检测到窃取隐私的恶意程序，如图 6-2 所示。

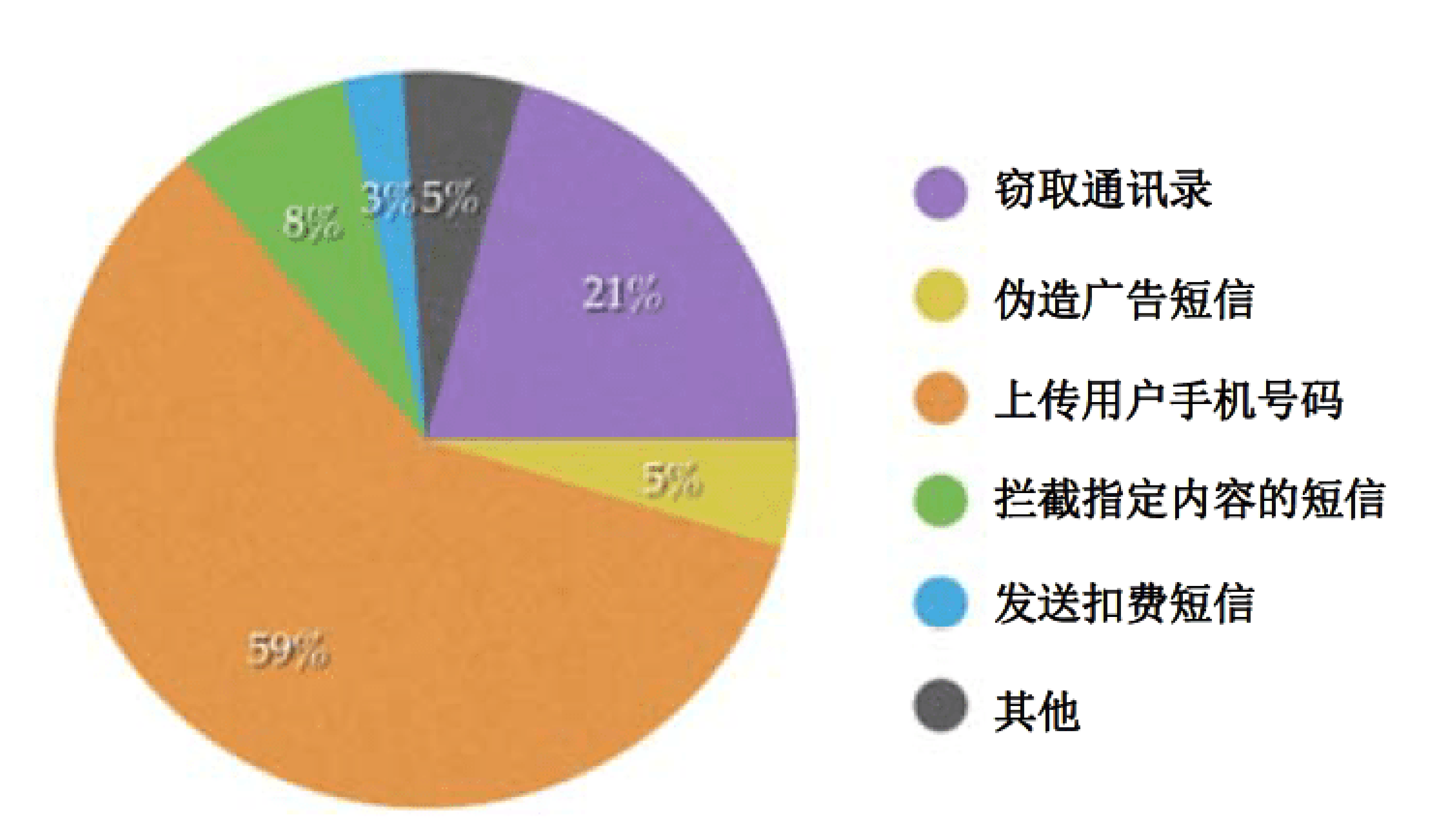


图 6-2　恶意窃取隐私行为

随着产生、存储、分析的数据量越来越大，隐私问题在未来的几年也将愈加凸显。所以，新的数据保护要求以及立法机构和监管部门的完善应当提上日程。

6.1.2　风险 2：数据管理困难

大数据除了有隐私方面的忧患外，它的危险还包括它将会诱使企业管理进入史诗般的同质性。收集足够的数据，每个人的统计开始看起来都是一样的。应用标准的分析，然后所有的结论也开始看起来都是一样的。正如营销人员们开始认为他们真正地知道他们所做的事情，但是他们会发现他们正在做的事情是其他人也正在做的。现在这不仅仅是没有创造力的问题了，而是积极地反创造力的问题。

无论从企业存储策略与环境来看，还是从数据与存储操作的角度来看，大数据带来的“管理风险”不仅日益突出，而且如果不能妥善解决，将肯定会造成“大数据就是大风险”的可怕后果。

事实上，很多企业并没有真正理解什么是大数据，也没有部署相关工具去有效地管

理它们。最近，LogLogic 与 IT 安全研究公司 Echelon One 共同完成了一项大数据管理调查，此次调查的对象是 207 位来自各行各业的主管或主管级别以上的个人，调查结果如图 6-3 所示。

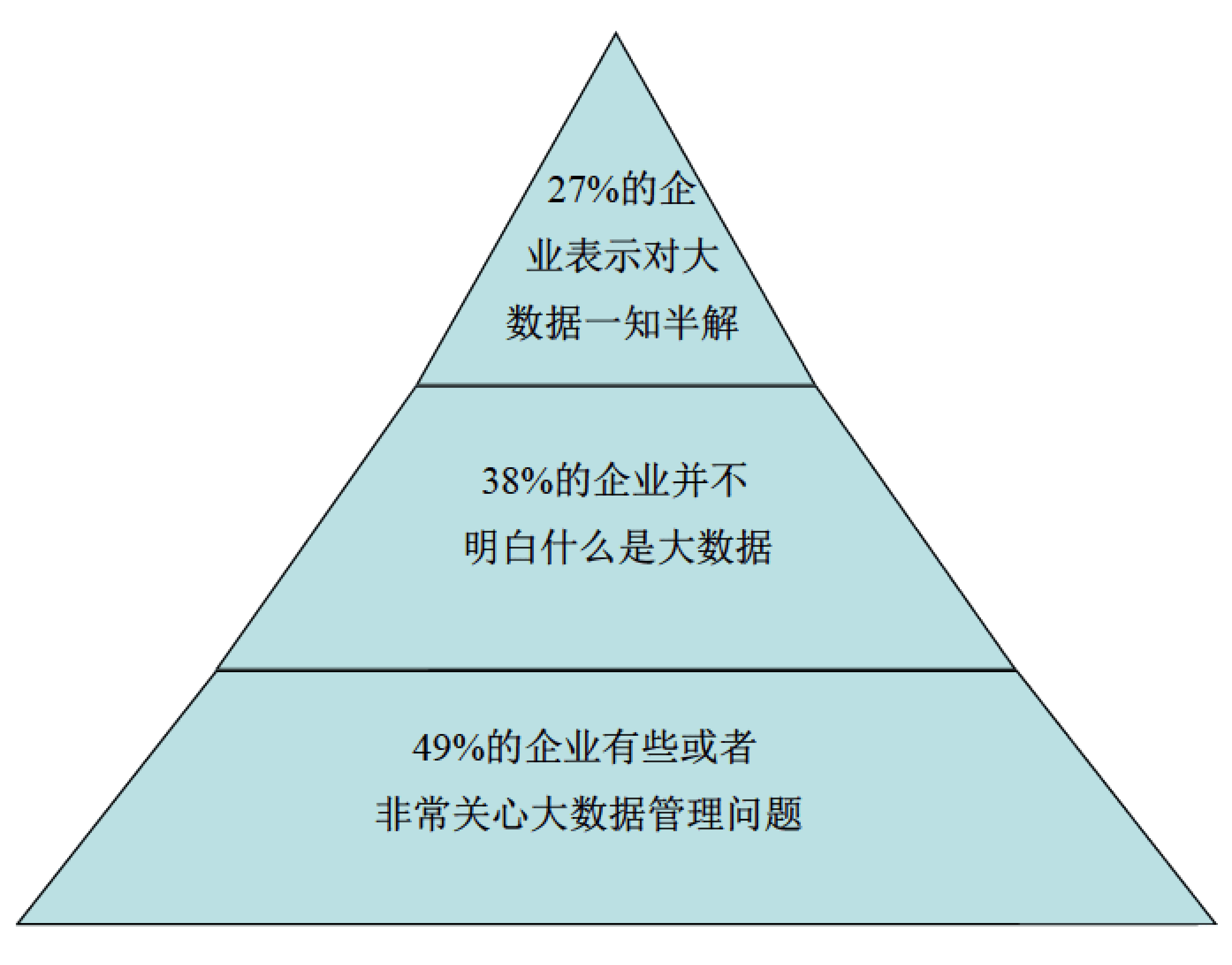


图 6-3　大数据管理调查结果

此外，调查还发现，59%的企业没有部署相关工具来管理 IT 系统中的数据，而是转向独立系统和其他系统，甚至使用电子表格。

如果正确使用大数据，它将为你提供梦寐以求的情报和洞察力，从而帮助企业做出明智的决定。在安全方面，它可以让你看到网络中正在发生的事情，以保护企业免受高持续性威胁和恶意软件。同时，它还能通过优化服务器和供应链管理来提高企业运营效率，甚至还可以帮助你处理法规遵从的问题。

专 家 提 醒

企业控制大数据的关键之一是日志管理，日志管理能够整合来自企业范围内的所有日志，建立索引存储库，并以常见的用户界面显示。因此，企业想要利用这些数据，就需要具备数据规范化和关联化以及报告和发送告警的能力。

6.1.3　风险 3：成本难以控制

随着时间的推移，企业产生的数据量已经越来越大了，这些数据包括客户购买偏好趋势、网站访问和习惯、客户审查数据等。传统的商业智能（BI）工具在处理企业海量数据时已经有点能力不够了。届时，你需要面对的是大量的支出：额外的人员和技术资源用以管理整体环境，例如系统管理及监控；通过不同业务系统而来的附加软件；以及管理集群的工具等。

例如，零售业巨头沃尔玛每小时处理超过一百万条客户交易，输入数据库中的数据预计超过 2.5PB——相当于美国国会图书馆书籍存量的 167 倍。通信系统制造商思科预计，到 2013 年互联网上流动的数据量每年将达到 667EB，数据增长的速度将持续超过承载其传送的网络发展速度。

另外，来自淘宝网的数据统计显示，淘宝一天内产生的数据量即可达到甚至超过 30TB，这仅仅是一家互联网公司一日之内的数据量，处理如此体量的数据，首先面临的就是技术方面的问题。海量的交易数据、交互数据使得大数据在规模和复杂程度上超出了常用技术按照合理的成本和时限抓取、存储及分析这些数据集的能力。

如图 6-4 所示，可以看出资源利用率低、扩展性差以及应用部署过于复杂是现今企业数据系统架构面临的主要问题。其实，大数据的基础架构首要考虑的就是前瞻性，随着数据的不断增长，用户需要从硬件、软件层面思考需要怎样地架构去实现它。因此，具备资源高利用率、高扩展性并对文件存储无障碍的文件系统必将是未来的发展趋势。

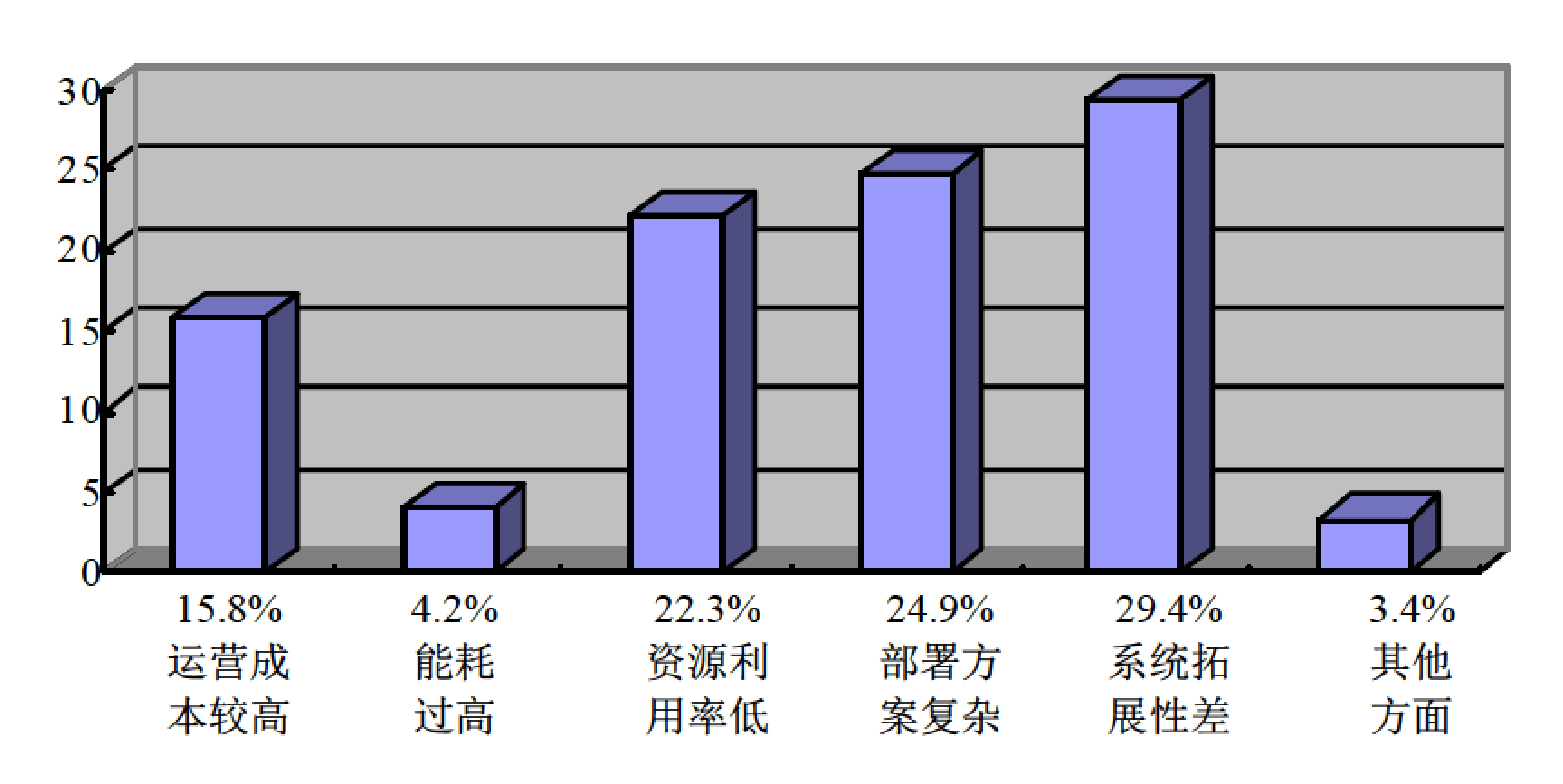


图 6-4　大数据构架面临的问题

由此可见，大数据对企业来说可能并不全是机遇，还意味着财政支出，原因是针对大数据存储或者挖掘的成本也很高。对此，笔者认为企业可以将重点放到通过最新收集的数据带来更多价值，减少非重点数据带来的存储硬件与软件的成本。

6.1.4　风险 4：网络安全漏洞

以前，只有 IT 部门那些最懂技术的工作人员才明白数据安全。在 IT 部门的办公室之外，病毒、木马、蠕虫这些词都不会被提及，管理层也并不关心黑客和僵尸机，董事会根本不清楚什么是零日攻击，更不用说零日攻击能带来多大的危害了。然而，现在，大数据以及随之而来的各种威胁几乎成为每一个单位日常的一部分，大数据的网络安全也慢慢地变成了一个被广泛关注的商业问题。

随着越来越多的交易、对话、互动在网上进行，这种刺激使得网络犯罪分子比以往

任何时候都要猖獗。影响和带来网络故障和安全事件的因素，主要来源于如图 6-5 所示的几个方面。

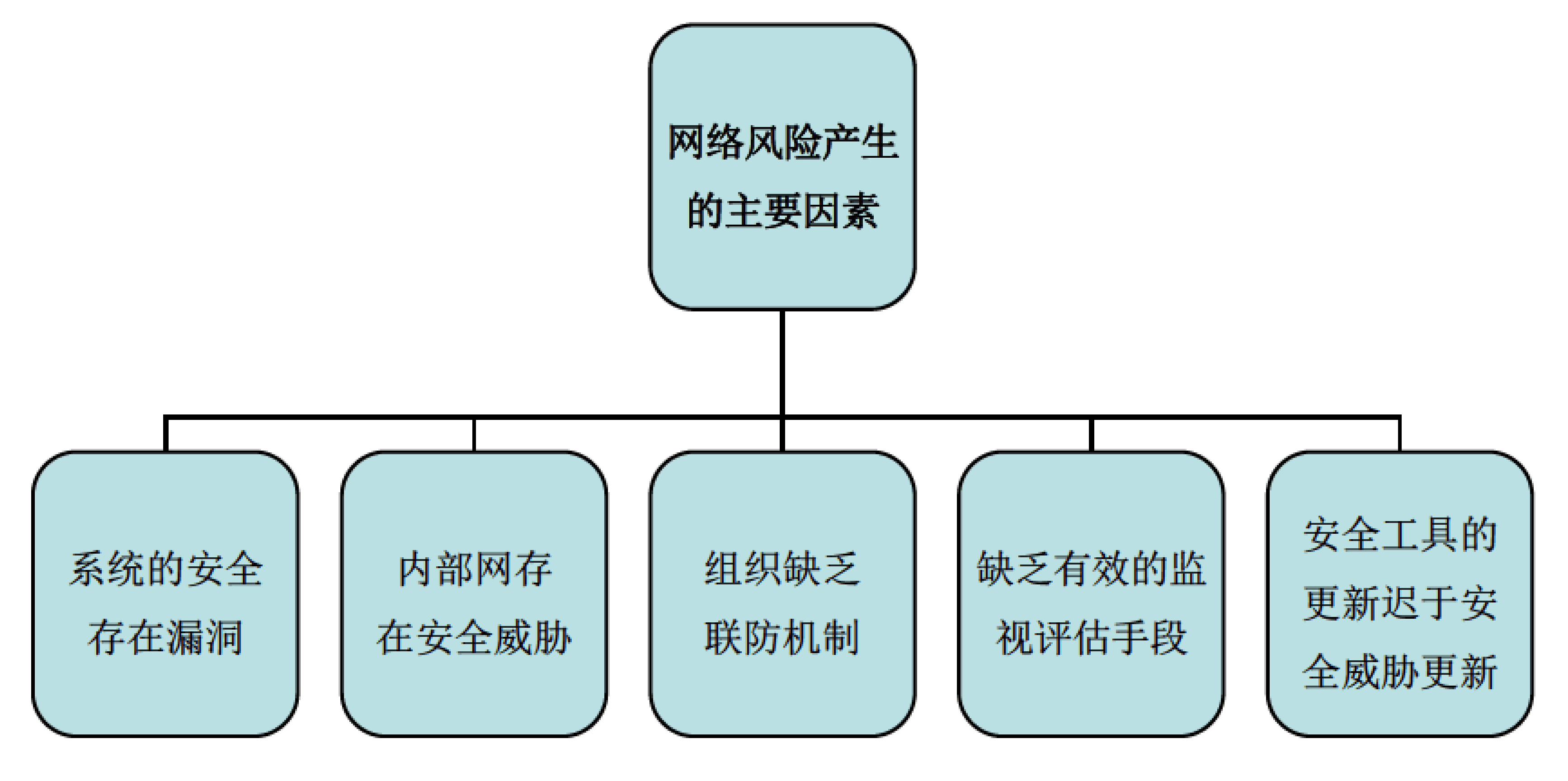


图 6-5　网络风险产生的主要因素

国际上，网络安全已开始从信息安全转向信息保障，从被动的预防向主动保护过渡。国内的信息保障虽已提上日程，但从理论走向应用还需要一个过程，这个过程的长短和企业信息化的进程息息相关。总的来说，网络安全系统以策略为核心，以管理为基础，以技术为实现手段。

专家提醒

很显然，保证数据输入以及大数据输出的安全性是个很艰巨的挑战，它不仅影响到潜在的商业活动和机会，而且有着深远的法律内涵。我们应该保持敏捷性并在问题出现前对监管规则作出适当的改变，而不是坐等问题的出现再亡羊补牢。

6.1.5　风险 5：数据人才缺乏

如今，大数据市场已经逐渐繁荣起来，但不少企业发现，目前对于最新的一些产品不能配备足够的人手。据塔塔咨询服务公司（TCS）的调查显示，IT 行业人才缺乏，符合条件的大数据分析人士很少，这也是许多企业在寻求打造与部署大数据系统所面临的困难之一。

如图 6-6 所示，在大数据时代，企业面临的挑战可以从中看出一些端倪。缺乏专业的大数据人才成为企业面临的最大挑战，其次是非结构化数据的分析和处理、传统技术难以处理大数据以及新技术门槛过高。

例如，阿里巴巴支付宝用户价值创新中心是支付宝大数据业务的核心部门，这个只有 7 个人的团队负责为公司开发出可以销售的商业化大数据产品。虽然阿里巴巴各类业务产生的数据为数据分析创造了非常好的基础条件，然而这个团队却因为招聘不到合适

的数据科学家而在研发上进展缓慢。

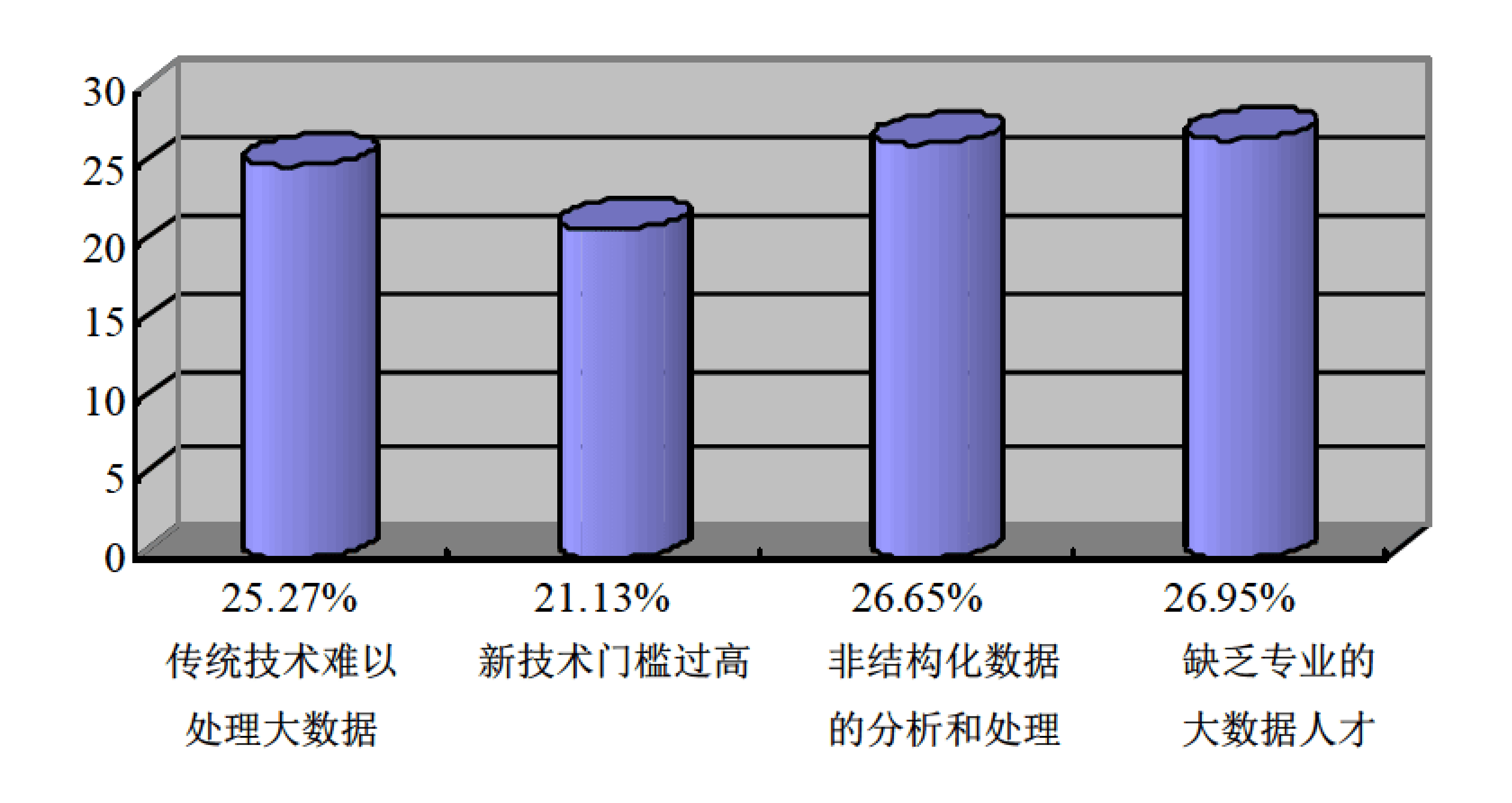


图 6-6 企业在大数据时代面临的挑战

不仅仅是阿里巴巴在面对大数据发展时遭遇人才瓶颈，多家咨询机构也都预测了大数据的快速增长和人才需求规模。据 Gartner 预测，到 2015 年，全球将新增 440 万个与大数据相关的工作岗位，且会有 25%的组织设立首席数据官职位。

在欧美国家，数据分析人员的工资水平可以排在前列，但国内数据分析人员整体逊于国外分析人员。笔者认为，大数据相关人才的欠缺将会成为影响我国大数据市场发展的一个重要因素。据 IDC 机构预测，中国大数据技术与服务市场将会从 2011 年的 7760 万美元快速增长到 2016 年的 6.16 亿美元。然而，国内各大公司普遍不重视数据分析人员，其薪酬水平属于平均水平。

大数据职位相关的技能主要包括数学、统计学、数据分析、商业分析和自然语言处理，数据科学家是复合型人才，需要对数学、统计学、机器学习等多方面知识综合掌控。目前，国内的人才市场上很难招募到优秀的数据分析人员。

因此，如果你正在寻找的是高端数据人才，这个任务无疑是很困难的。不过在你发出“我找不到人才”这样的歇斯底里之前，确定好你的需求和培训的规模，然后和当地一所大学建立联系，这样或许你的问题会变得更容易解决。

6.2 步步小心，大数据项目 7 大误区

大数据分析可以给组织带来很大的商业价值，但是如果你不小心，不从其他公司犯的错误中吸取教训的话，它也可以带来灾难。因此，应谨记本节提到的几个问题，切莫成为大数据分析项目的反面典型。

6.2.1 误区 1：盲目跟风

由于“大数据（Big Data）”近两年来是信息技术领域最时髦的词汇，因此，很多人甚至还没明白什么是大数据，就眼高手低地开始部署大数据项目，妄图赶上大企业的步伐，想走捷径，结果往往是钻入了“牛角尖”。

很多企业或机构在开发他们的第一套数据仓库或者 BI 系统时经常会犯“盲目跟风”的错误。太多时候，大数据分析项目管理者被技术炒作所迷惑，忘记了他们首要的任务是商业价值，过分追求数据分析技术，却不知那仅仅是一个用来产生商业价值的工具。

现在应对大数据，可以以高可用高可靠性、高可扩展性的基础架构和高性能的分析系统来应对，然而，谈大数据的风险，谈数据挖掘，它的效果到底多好？事实上是需要得到验证的。

笔者认为，尽管大数据是个值得重视和关注的方向，但目前技术上并不成熟，各企业不要盲目上马大数据项目、建大数据中心，以免重蹈云计算过热的覆辙。另外，云计算发展几年来成效并不显著，很多地方建的云计算中心利用率不高，不少还仅仅是数据库，没有提供云服务的能力。

大数据分析的支持者们不应该盲目地采用产品，他们首先需要判断该技术所服务的业务目标，以便建立业务案例，然后为手头工作选择正确的大数据分析工具。如果没有对业务需求的深刻理解，会存在很大风险，项目团队最终可能将创建出一个毫无用处的“大硬盘”。

因此，规避大数据的风险，不能盲目跟风，特别要明确实施大数据的目标，要有切实可行的规划，此外要有质量足够好的数据。尤其是发展大数据产业需要有明晰的产业规划，建大数据中心要有明确的用途和服务对象。

专家提醒

笔者再次提醒，大数据时代确实给我们带来了很大的诱惑，我们可以通过数据分析得到预知未来甚至穿越过去的效果，但是我们也不要盲目跟风，适合自己的才是最好的。

6.2.2 误区 2：思路太过僵硬

很多情况下，企业的大数据项目采用“放羊式”管理：寻找到一片草地，就把羊赶出去，任羊自己去寻找水源和青草。结果往往是：聪明的羊骠肥身圆，迟钝点的羊瘦骨伶仃。这是由于万物生存法则——“适者生存”所导致的。

通常，人们总是不断尝试他们过去的做法，即便当他们面对不同的场景时也会这样。从而导致在大数据情况下，一些企业会想当然地认为所谓“大”只是意味着更多的交易和更大的数据量。这种观点可能是正确的，但是许多大数据分析策略会涉及非结构化和

半结构化信息，需要以完全不同于企业应用程序和数据仓库中结构化数据的方式管理和分析。

因此，企业管理者不仅要让“大数据正确地做事”，更需要“引导大数据做正确的事”，最好有一套新的方法和工具来进行大数据的捕获、清洗、存储、集成和访问。正如一个好棋手，走一观二想三，深谋远虑才能保证在大数据道路上不断前进。

专家提醒

创新性思维为我们提供了科学的思维依据和方法，将其融会贯通后定会提高大数据分析问题的能力和解决问题的能力，促进企业快速发展。

6.2.3 误区 3：不注重他人的经验

在做大数据项目时，有些人会走向另一个极端，认为大数据中的一切都是完全不同的，他们必须从头开始，从而不知不觉地走进了误区。对于大数据分析项目的成功，这种错误甚至比认为没有不同更要命。

俗话说：“失败是成功之母。”每个人都熟悉的这句话，同样可以运用于大数据项目。其实，数据分析大师是经过无数次失败才换来成功的。因此，各企业的大数据项目往往只是分析的数据结构不同，而数据管理的基本原则却都大同小异，完全可以借用，这样才能更节省时间和精力。

6.2.4 误区 4：把大数据当“门面”

现实中，有些企业喜欢追求热门，只是将大数据项目当作“噱头”来吸引业务，认为自己有了大数据项目就是新型科技企业，却不看重大数据的实际价值。据国外报告显示，多数企业只用了收集到的数据总量的 0.5%来进行决策，这意味着绝大多数的数据被浪费掉了。

在这些企业中，衡量大数据分析项目的成功仅仅是通过数据收集和分析来进行。而事实上，收集和分析数据只是开始。如果结合了业务流程，并促使业务经理们和用户为改善组织绩效和业绩而付诸行动，之后，分析才能产生商业价值。要获得真正的效率，就需要把分析项目纳入反馈闭环，以便于交流分析结果，然后基于经营业绩提炼分析模型。

大数据的应用不仅仅停留在 IT 领域，在医药、科学、制造以及气象等行业，都将出现海量的数据应用，如果能合理地利用这些资源，其将对行业产生巨大的推动，但目前来看，大数据应用还远远不够。多数企业仍然是扔掉的数据比保留的多，如何去筛选数据，数据留存多久，这一系列问题都是需要企业与监管部门面对的，但现在仍然缺少一个大数据应用的框架。

6.2.5 误区5：过度夸大数据成果

近日，笔者听到两个朋友抱怨。

朋友A说："我们的领导不干脆。外部门踢过来的工作，不说接也不说不接，搞得下面的人做也不是不做也不是。对下属的求助也是模棱两可，总是说，'这个事儿，再搞搞，再看看，再研究研究，'很多都明确了的事儿还是要一拖再拖，不决策。"

朋友B说："我们的领导不懂业务，又喜欢揽活，经常胸脯一拍说，'这个事儿我来干！'回来就丢给下面的人做。但是实际上这个活儿与我们部门是'风马牛不相及'，根本就无法完成，强出头的结果往往是费力不讨好。"

这样的对话每天都在发生，这样的领导也比比皆是。不承诺和过度承诺，已经成为管理者们常见的一个现象。究其根源，往往是不了解业务、流程及对责任感的错误理解所致。其实，许多大数据分析项目陷入了这样的一个误区：过度宣扬他们部署的大数据系统会有多么快，业务会获得多么重大的益处。

企业对大数据项目的"过度承诺"需要在销售过程中向客户明示，这种"过度承诺"在客观上使该项目成为"卖点"，刺激了客户购买欲，增加了相关的商品销量和扩大了营业额。但是，长此以往，结果往往却不乐观。过度的承诺和交付的不足，必然导致业务与技术的分离，造成该组织会在很长时间内推迟特定技术的选用——即便其他许多公司已经使用该技术获得了成功。此外，如果你设定了很轻松、很快就能获益的预期，业务主管就有一种认识倾向，容易低估了需要参与和承担义务的程度，当足够资源不能兑现时，预期的收益就很难达到了，那么你的大数据项目基本就贴上了"失败的标签"，甚至还要承担客户的损失。

6.2.6 误区6：想要获得所有数据

我们正生活在一个前所未有的大数据时代当中，我们从来都没有像现在这样能够获得如此多的数据。在如今的工业化社会中，平均每个人一天所消费的信息量超过了生活在十五世纪的人一生所消费的信息量。

很多企业为了挖掘大数据，不断地构建、升级自己的IT系统，妄图获得所有的数据。其实，目前还没有一个人或一家公司能够存储和检索关于某一特定主题的全部数据，更不要说是所有数据了，包括谷歌在内。谷歌索引的只是表层网中的信息，而不是深层网中的信息。专家估测，后者的规模是前者的25倍。因此，在我们进行搜索时，我们所获得的信息量仅仅是互联网信息量中的4%～6%。

笔者认为，钱必须要用才有价值，数据也是一样。只有不停地使用数据，挖掘数据背后的关系和价值，才能如滚雪球一般，使数据之间的相互关系更丰富和完善。

6.2.7　误区 7：认为软件是万能的

很多人构建一个大数据项目，是希望他们部署的软件会神奇地实现一切功能，把所有的问题都丢给分析软件，不再愿意亲自去动脑思考。当然，人们应该明白希望总是比现实更美好。软件确实会带来帮助，有时帮助还会很大，但是大数据分析的效果取决于被分析的数据和使用工具的分析技能。

大数据在某种意义上只能作为一个工具，不能代替人类自己的分析，如果把所有的事情都交给大数据来处理很可能就会陷入一个非常大的困境。例如，现在很多影视公司在制作影视作品时，通过大量的数据分析来指导创作，这看起来似乎是合理的，但是实践结果往往并非如此。国内一家知名的影视数据分析公司的影视剧都是在海量的数据分析基础之上进行创作的，包括什么样的题材、什么样的演员、什么时间投放都经过了非常精密的计算，可是最终理性地看市场效应，在业内有影响力的作品并不多。

由此可见，在应用数据软件指导商业行为的时候，依然存在着很多不确定性。这就需要大家回过头来思考另外一个问题，即大数据对商业行为的产生或产生的影响体现在什么地方。笔者认为其更多是在营销领域，通过一个软件分析消费者的主要需求，然后根据需求选择相应的商品进行生产。同时，也可以根据消费者的需求对已有的商品进行修改完善。所以，从这个意义上讲，大数据对各个领域的影响肯定是巨大的，如果能够很好地运用，对于企业的发展有非常大的作用，但是过于迷信也可能会变成谬误。

专家提醒

当然，笔者并不是说，因为存在不确定性，大数据就不能为我们提供帮助了，不能将减少不确定性和消除不确定性混为一谈。大数据能够帮助我们消除不确定性的这一天还没有到来，可能这一天永远也不会到来。对海量非结构化数据进行分析或许能够帮助公司更好地理解客户的情绪，但不要误认为大数据能够为我们排除所有的可能性，生命的无常和业务的起伏将会破坏我们制订出的完美计划。

6.3　踏雪无痕，彻底逃离大数据监视

美国作家艾伯特-拉斯洛·巴拉巴西的新书《爆发》中有一个这样的片段：“我点击了自己的名字，页面上出现了一张熟悉的照片——是我穿着一件蓝色衬衫的照片，旁边配有我的基本履历资料……我点开了一个最近更新的链接，地点是波士顿的马萨诸塞大街……两秒钟后，我在视频中看到了自己推开了地铁站那厚重的大门……每次看到自己出现在视频中，我都会浑身不自在。但现在可好，我的一举一动已经被 LifeLinear 网的系统给记录了下来……”

书中的“LifeLinear 系统”只是作者杜撰出来的，并非真实存在。但是作者同时认为，在科技发达的今天，借助大数据的平台，“LifeLinear 系统”并非不能实现。这样的场景又让人毛骨悚然，如果真有这样一套系统面世，我们的隐私岂不是要暴露在光天化日之下？

大数据堪称一把双刃剑，不论是企业还是个人，都会因大数据的爆发获益匪浅，但同时个人隐私也无处遁形。如今，人们在网上的每一次活动，都会留下蛛丝马迹。虽然我们无法完全躲避“大数据”的监视，却也可以踏雪无痕、隐遁无形，逃离那些秘密网络跟踪。

6.3.1 码头：让网络行为一目了然

“码头”项目及其监控手段是 NSA（美国国家安全局）所实行的监控项目中最鲜为人知的一个，即使是那些参与其中的情报专家对项目整体也知之甚少。“码头”项目所监控的电子邮件、网上聊天系统以及其他借助互联网交流的媒介使用频率在当下远胜于普通的电话或者手机。

美国在 2001 年“9・11”恐怖袭击发生后不久启动了“元数据”项目，NSA 将这些“元数据”视为“数字网络信息”。这一项目收集互联网“交通”原始数据，被称作“码头”项目，也称为“大块互联网元数据”项目，其包含互联网信息发送双方的地址，包括可以显示发送或接受信息者所在确切位置的 IP 地址。项目启动之初以一方为美国境外的人或外国人之间的通信为限定范围，但 2007 年拓宽至美国公民以及居民。

“码头”项目的 IP 记录功能就像是一个导航记录，你曾经看过的内容，曾经在网上发过的帖子等，只要它了解你的 IP 记录，它就像在看日记一样地了解你的行为。对于这些信息的追踪及分析，能切实知晓一个普通的美国民众是否与一个臭名昭著的恐怖分子有所联系。同样，基于这些信息，个体的健康状况、政治或者宗教信仰、涉密的商业谈判，甚至是否存在婚外情等状况，都能一目了然。而这恰恰是美国民众最为担心的，也是美国政府极力回避的所在。

6.3.2 上游：截取全球互联网数据

与“码头”类似的监视项目，还有“上游”（Upstream）项目，其通过美国周边的海底光缆搜集情报，截取全球互联网数据。美国《华盛顿邮报》2013 年 7 月 10 日公布了一张美国国家安全局的机密幻灯片，其中对“棱镜”计划以及与之平行展开的“上游”计划有所介绍，如图 6-7 所示。

在这一张最新公布的演示图中，上半部分蓝色框内是“上游”项目，显示了从美国东西海岸延伸至世界各地的深海光缆路线，意思是从海底光缆等基础设施截取数据。海

底光缆对世界范围内的数据传播极为重要，对美国及其盟友的监控项目也有举足轻重的影响。

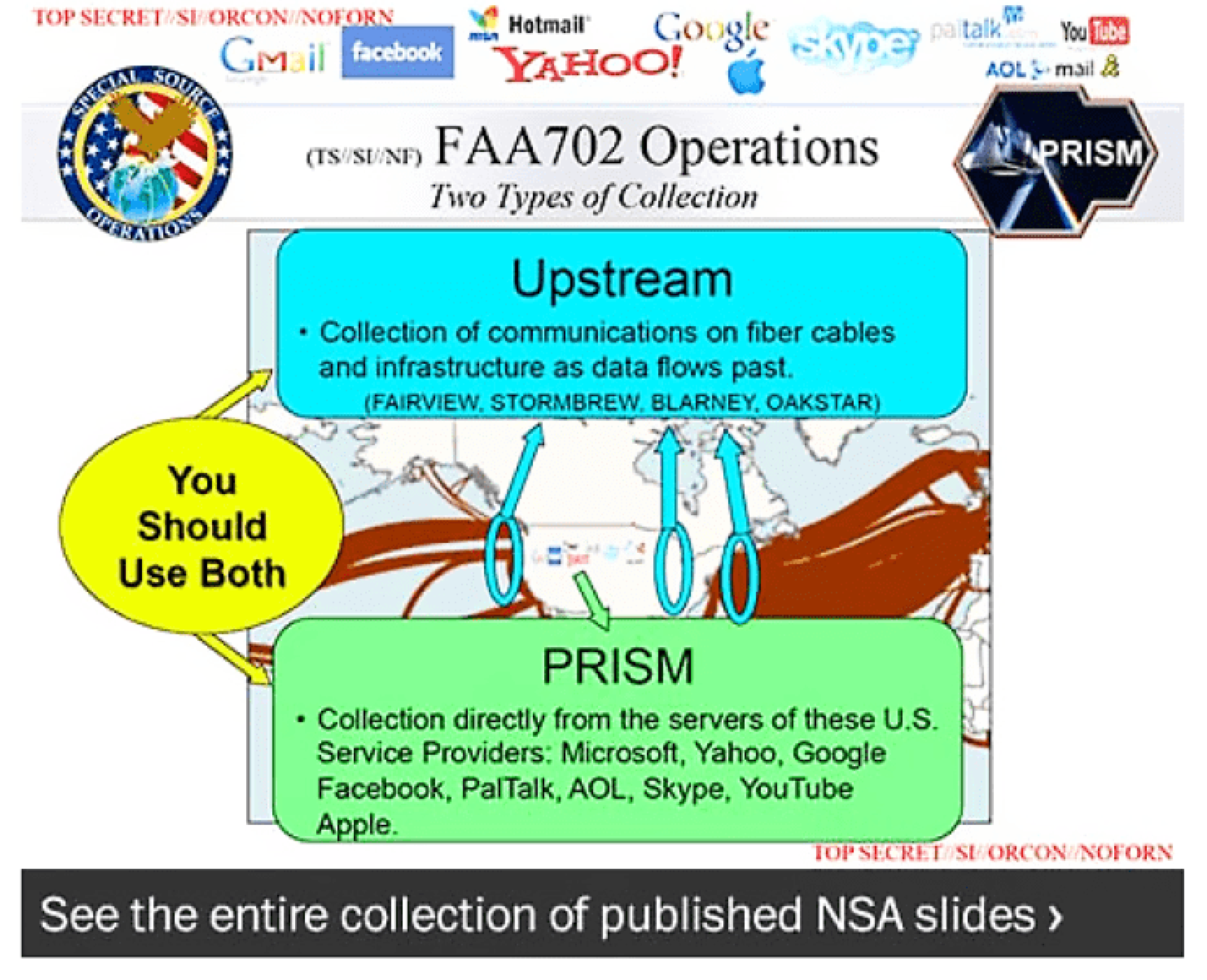


图 6-7 美国"上游"监视项目

图片下半部分绿色框内解释了"棱镜"计划，它通过谷歌、微软、脸谱、雅虎、Skype、PalTalk、Youtube、苹果和美国在线等 9 家互联网企业挖掘数据，其中的文字介绍是"直接从服务器上搜集信息"。

幻灯片还用黄色圆圈提醒国家安全局人员"应利用两个项目"。为保障"上游"项目的顺利实施，美国国家安全局和国防部等机构在 2003 年与美国环球电讯公司签署《网络安全协议》。据悉，环球电讯公司的海底光缆覆盖全球四大洲的 27 个国家和地区。在过去 10 年中，有更多的电讯公司签署了类似合作协议。

专家提醒

每个人都期待获得个性化服务。但是，在大数据时代，想要获得个性化服务，就一定会在某种程度上牺牲自己的隐私。

6.3.3 棱镜：备份全球互联网数据

美剧《疑犯追踪》里有这么一件"神器"：它几乎无所不能，全天候监视所有人的行踪，聪明地预测出谁是危险分子，谁会遭遇不测……美国政府用它攻击恐怖分子，开发者则用它拯救普通人。这不只是一部科幻剧，它也出现在现实的世界里，即美国的"棱镜"项目。

首先让我们回顾下轰动 2013 年的“棱镜门”事件。2013 年 6 月，美国前中情局（CIA）职员爱德华·斯诺登将两份绝密资料交给英国《卫报》和美国《华盛顿邮报》，并告之媒体何时发表。按照设定的计划，2013 年 6 月 5 日，英国《卫报》先扔出了第一颗舆论炸弹，即美国国家安全局有一项代号为“棱镜”的秘密项目，要求电信巨头威瑞森公司必须每天上交数百万用户的通话记录。2013 年 6 月 6 日，美国《华盛顿邮报》披露称，过去 6 年间，美国国家安全局和联邦调查局通过进入微软、谷歌、苹果、雅虎等 9 大网络巨头的服务器，监控美国公民的电子邮件、聊天记录、视频及照片等秘密资料。

“棱镜”计划是“上游”项目的兄弟，相当于“下游”项目，其收集的是经过科技公司加工的数据。根据报道，代号为“棱镜”的监视项目从 2007 年开始实施，从未对外公开过。接入互联网公司的中心服务器可以让情报分析人员直接接触到所有用户的数据，通过音频、视频、照片、电邮、文件和连接日志等信息，跟踪互联网使用者的一举一动，以及他们的所有联系人，如图 6-8 所示。



图 6-8　“棱镜”监视的网络信息类型

专家提醒

从技术角度看，棱镜是正宗的大数据武器。虽然还不如《疑犯追踪》里的机器万能，但足以让大家担心个体隐私不保。人们更害怕政府拥有大数据后，权力和能力膨胀，必然滋生腐败。数据如万川归海般途经美国，“山姆大叔”便可架网捞鱼，坐收渔利。棱镜数据监测的原理也是如此，就像三棱镜把自然光分成红、橙、黄、绿、蓝、靛、紫七色，在光纤上接入“棱镜”，可以让光纤传输的信号一览无余，通过大数据系统进行分析挖掘。

在过去 6 年中，“棱镜”项目经历了爆炸性增长，眼下美国国家安全局约七分之一的情报报告依靠这一项目提供原始数据。可以说，“棱镜”项目以近乎实时备份的方式，

备份了整个全球互联网的全部数据。利用这些备份数据，可以拼出一个人一生的网络足迹。

由此可见，因为具备足够资金、技术和不受限的权力，政府机构等大组织是大数据的最大受益者，可肆意窥探个体的网络活动和关联网络。不过，现有的大数据技术，擅长利用历史记录来预测已有事物在未来是否出现，并不擅长判断从来没有先例的事物。要防范大数据技术滥觞，需要发挥个体的创造性，不要成为机器眼里可以预测的循规蹈矩者。

6.3.4 星风：监视全球通信大数据

斯诺登揭开的“棱镜”项目只是美国政府秘密监视系统的“冰山一角”。据《华盛顿邮报》爆料称，斯诺登曝光的“棱镜”项目，源自此前从未公开的“星风”（STELLARWIND）秘密监视计划。

“星风”计划成立于 2004 年，不过由于当时的法律程序等敏感问题，时任小布什政府被迫做出让步，缩减在美国本土的监听项目。与此同时，为了避免“星风”计划的夭折，小布什政府将其拆分为“棱镜”（PRISM）、“主干道”（MAINWA Y）、“码头”（MARINA）以及“核子”（NUCLEON）4 大项目，均交由美国国家安全局（NSA）执掌，如图 6-9 所示。

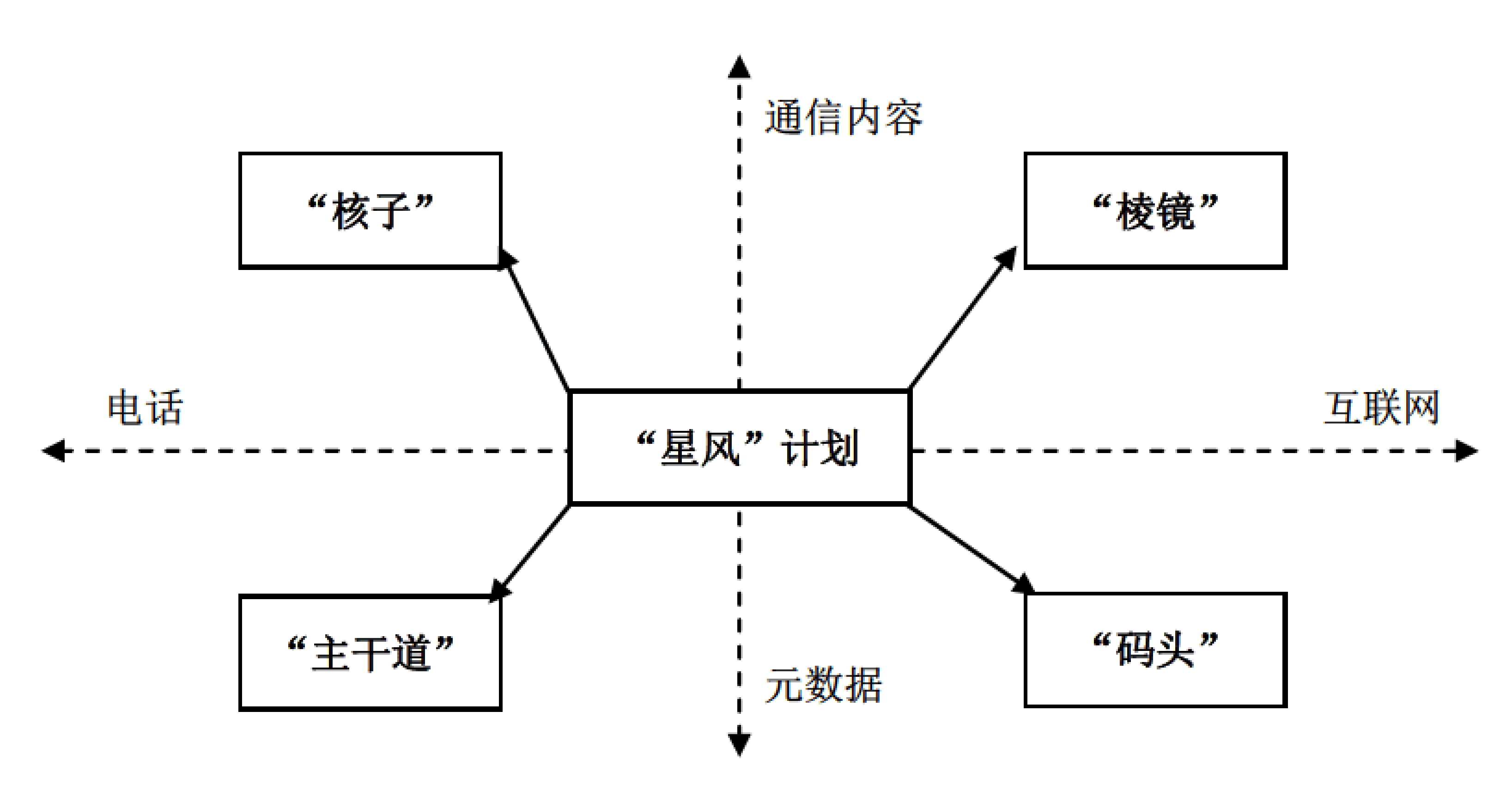


图 6-9 “星风”计划的主要内容

时至今日，“星风”计划对于很多美国人来说是待解之谜，而唯一能大致确认的则是由“星风”计划拆分出的 4 个监视项目，它成功帮助小布什和奥巴马政府对全球范围内的现代通信数据实行了有效监控。

《华盛顿邮报》表示，“主干道”和“码头”秘密监视项目分别对通信和互联网上数以亿兆计的“元数据”进行存储和分析。“主干道”项目负责秘密监视电话信息，包括通话或通信的时间、地点、使用设备、参与者，但不会窃听通话内容。从 2009 年一

份流出的机密材料来看，美国国安局花费了 1.46 亿美元的反恐基金购买硬盘等设备，用于存储“主干道”秘密监视项目上的元数据。另外两个“规模小得多”的“棱镜”和“核子”秘密监视项目则负责截取内容。其中，用来截获电话通话内容及关键词的叫“核子”秘密项目。

尽管按照美国情报部门的说法，这些秘密监视项目的目标都是“外国人”，但事实上，四大情报搜集计划牵涉的范围极为广泛，从某种程度上说，几乎可触及每一个美国家庭。

专家提醒

元数据（Metadata）是指在地理空间信息中用于描述地理数据集的内容、质量、表示方式、空间参考、管理方式以及数据集的其他特征的数据，它是实现地理空间信息共享的核心标准之一。例如，在对电话和互联网监视的语义下，元数据主要指通话或通信的时间、地点、使用设备、参与者等，不包括电话或邮件的内容。在美国，法律对于元数据的保护很少。而根据新技术，监视机构有效挖掘元数据的能力，已经比窃听和截取通信内容更加重要。

6.3.5 小甜饼：窃取个人网络隐私

2014 年新年即将到来，笔者好友张莉经常浏览汽车网站，准备买台新车回老家过年。不久，张莉便发现，在看了几个汽车网站后，即便是在与汽车无关的页面，也看到了比过去更多的汽车广告。这就是 Cookies 在“作怪”，电脑中的 Cookies 记录了张莉对汽车的兴趣，便向她推送相关的广告。

“通过 Cookies，我们什么都能知道，包括你的性别、年龄、职业、收入。”2013 年央视“3·15 晚会”上这段关于 Cookies 泄露个人隐私的视频，让原本“默默无闻”的 IT 术语一夜之间红遍了全国。

Cookies（昵称为“小甜饼”）也被称为 HTTP Cookies、网络 Cookies 或浏览器 Cookies，它是当用户浏览网页时，网络服务器以文本格式存储在用户电脑硬盘上的少量数据。Cookies 的主要目的在于帮助网站记忆用户之前可能进行的操作，自 1993 年问世至今已经过去了整整 20 年。

对普通用户来说，Cookies 主要用来判定注册用户是否已经登录网站，这样可以免去用户重复登录网站的麻烦，试想如果你刷新一次微博都需要重新登录，想必就没有多少人愿意上网了。Cookies 的另外用途是网上购物的“购物车”功能。用户可能会在一段时间内在同一家网站的不同页面中选择不同的商品，这些信息都会写入 Cookies 以方便最后网购结账。

但是，某些第三方广告公司往往通过采取在网站加代码的方式窃取用户的 Cookies，这些 Cookies 几乎覆盖了所有网民群体，并通过分析 Cookies 来收集用户的 IP 地址、

账号、身份、联系方式等信息，用于广告营销，但这显然没有充分尊重用户的知情权和选择权。

Cookies 的存在最初是为了方便用户使用，然而被一些有商业企图的机构在用户并不知情的情况下，采集并加以商业运作，那就是不折不扣的违法行为，正是这种“网络臭虫”的存在，让 Cookies 有了隐患，危及到用户的隐私安全。

“网络臭虫”通过在用户广泛访问的网页上放置一个像素大小的图片（代码），而用户根本看不到这张图片。“网络臭虫”的工作就是通过获取 Cookies 来获知用户的浏览习惯，进行隐蔽的跨网站跟踪行为。这个页面一天内如果有 1000 万人访问，那么该公司一天就获取了 1000 万份个人信息。更可怕的是，网络黑客可以通过木马病毒盗取用户的 Cookies，直接骗取网站信任，无需输入用户的账号和密码即可登录网站。

针对这个问题，微软公司最新的 IE 10 浏览器中默认开启 DNT（Do Not Track，直译也就是“不追踪”）“禁止跟踪”功能。另外，国内的 360 安全浏览器都推出了“禁止跟踪”功能，可以有效阻止某些网站的 Cookies 跟踪和跨站跟踪行为，对于那些不遵守禁止跟踪协议的网站，许多浏览器还提供了隐私保护浏览器模式以及 Cookies 清理功能。同时，许多浏览器软件推出的多项清理功能，也无疑给用户提供了自主保护个人隐私的工具。如图 6-10 所示为搜狗浏览器的“清除浏览记录”对话框。

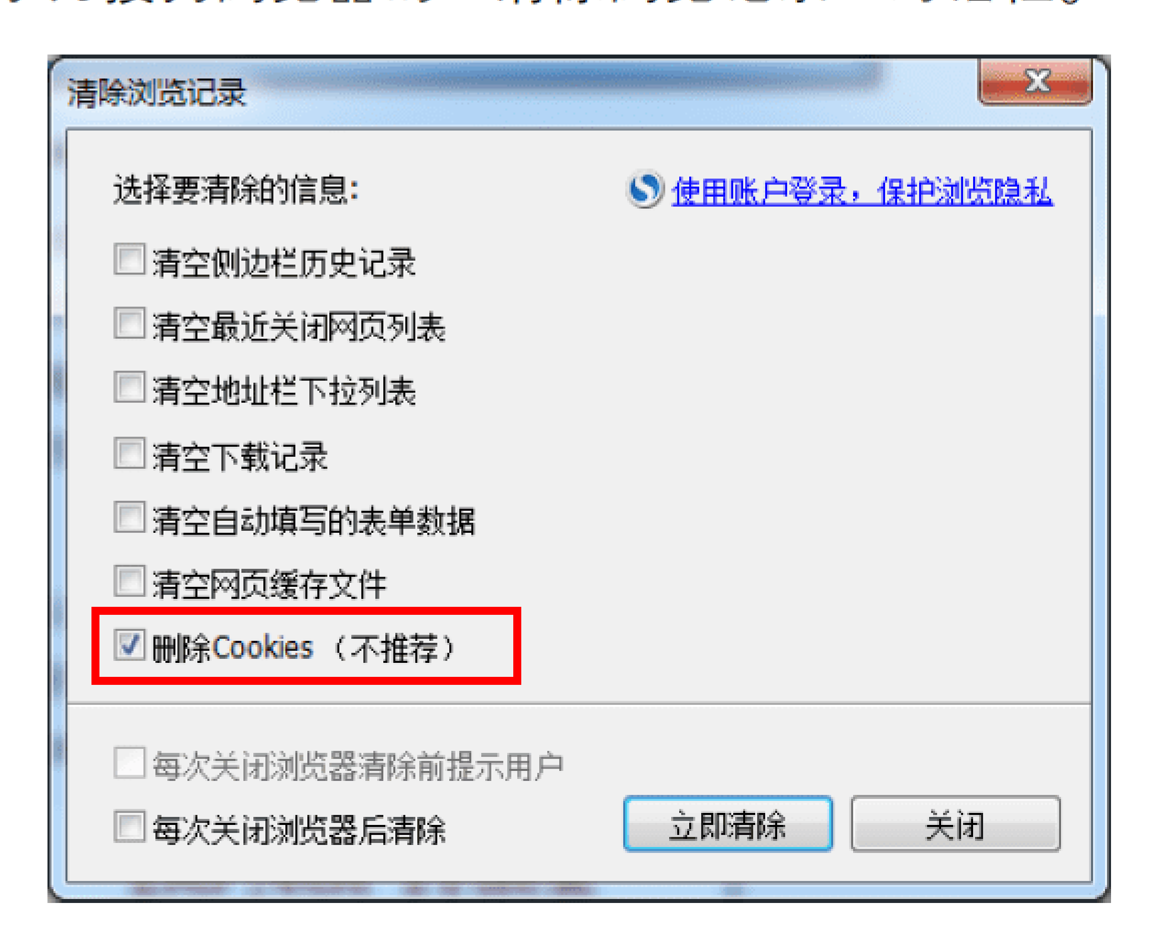


图 6-10 搜狗浏览器的“清除浏览记录”对话框

尽管如此，笔者建议用户还应从自身做起，不要在不清楚来源的网页上填写任何个人信息，例如你的年龄、性别、收入等，你在不同网站填写的信息很可能会被其他人获取后整合得到你的全部信息。

6.3.6 间谍软件：让我们无处藏身

在大数据时代，聪明人已经极端地依靠互联网来达到各种目的，其中最重要的就是：

发现用户，研究用户，最终控制用户。互联网之父，英国南安普敦大学的计算机科学教授伯纳斯·李曾经说过："我很担心通过搜集在线数据描绘网络用户特征和详细了解用户的习惯。避免这种窥探行为是非常重要的。"

随着网民数量的急剧增长和移动网络的普及，网民在电脑或手机设备上存储的账号、密码等机密信息也越来越多，以窃取用户机密文件和个人隐私为目的的"间谍"软件已经超过传统意义上的病毒成为网民的最大威胁。

"间谍软件"是一个概括性的术语，用来描述通常未事先适当征求用户同意便执行某些行为的软件。间谍软件能够在用户不知情的情况下，在其电脑上安装"后门"，搜集、使用并散播用户的个人信息或敏感信息，如图 6-11 所示。



图 6-11　间谍软件的作用

据悉，英国网络安全公司 ScanSafe 近期推出一项新型的间谍软件屏蔽管理服务，在对该软件进行的 10 周示范运行时，公司发现从受感染计算机发出的间谍软件通信流量能占到总网络流出流量的 8%。此外，间谍软件现在变得越来越狡猾了，它们把其外出流量夹杂在正常的网络流量之中。对于电脑用户来讲，感染上这些间谍软件会导致他们电脑中的私人信息失窃。

ScanSafe 公司称，目前间谍软件共占网络盗窃事件的 20%，目前还有增长的趋势。一些恶意程序如 CoolWebSearch 现在采用新开发的 root-kit 结构，可以躲过杀毒扫描。

对付间谍软件是一场永远不可能结束的战争。这已经成为现代计算环境中一道"亮丽"的风景线。而且像所有的战争一样，与间谍软件的战争也涉及防御和进攻的策略问题。正确运用下面的一些技巧可以帮助你免受恶意程序设计人员和黑客的危害。

- 防火墙：防火墙就像站在你的计算机或私有网络门口的一位"警卫员"，它会阻止进入或发出的不符合设定标准的数据通信。
- 反间谍软件：主要用于搜出计算机内隐藏的间谍软件、特洛伊木马、蠕虫等，是

迎战黑客和间谍程序的有利武器。同时，要保证你的反间谍软件程序拥有自动更新特性。

- 查看邮件要小心：在多数情况下，查看电子邮件需要格外当心。最起码不要打开来自并不认识的人或组织的附件，还要提防那些“道貌岸然”的像是来自某个官方网站的邮件，它们可能向你索要关键信息。
- 正常关机：为了保护你自己，在不想用电脑时可将其关闭。如果你实在不愿意关闭电源，可以在不使用网络时，通过防火墙或其他方式关闭网络连接。

6.4 有备无患，做好大数据风险管理

避免大数据的管理风险的第一要务，并非是技术或产品上的实施与部署，最重要的应该是策略与理念上的转变：大数据首先不是机遇而是挑战，首先需要着手解决的不是数据分析、利用，而是将数据更好地存储与管理起来，这才是大数据时代首先要做的事情。

6.4.1 风险管理利器 1：IBM StorWize V7000

在数据管理时，将所有数据放在一个地方是有很大风险的，为了数据的安全，数据应该存储于不同的地方。如数值数据可以存储在数据库里，非结构化的数据则可以存储在文档或者表格里。这样将风险信息可能的来源进行了细分，意味着我们可以迅速了解综合风险状况。

在如今的存储管理环境下，打破复杂性升高和数据爆炸式增长的循环可能是一大挑战，购买和管理存储设备的老办法已变得不那么有效。IBM StorWize V7000 是 IBM 最新发布的一款中端存储产品，在发布这款新产品之前，IBM 特意为其制作了具有强烈神秘感的广告，并宣称这将是“改变存储游戏规则”的产品，如图 6-12 所示。



图 6-12 IBM StorWize V7000

确实，IBM 一直是主打性能稳定的招牌，其中这款 IBM StorWize V7000 作为目前热卖的磁盘列阵，它可充分保护企业的数据安全，该机支持 12 块 3.5 英寸磁盘驱动器，用户在不中断系统运行的情况下，可以将数据迁出现有存储设备，从而简化实施流程并

且可最大限度地避免用户服务中断。

IBM StorWize V7000 为用户提供了与虚拟化服务器环境互为补充的虚拟化存储系统，其具有无与伦比的性能、可用性、先进的功能和高度可扩展的容量。配置的方面，IBM StorWize V7000 高速缓存达到 8GB，每个机柜可以组合 12 个 SAS 驱动器，支持 RAID 0、1、5、6 和 10 接口，并且硬盘转速达到 10000rpm、近线 7200rpm，可谓是性能强悍。

通常情况下，在多套存储系统中，统一执行存储层的数据灾备可以说是难上加难的工作，不仅需要分别购置和部署每套存储上的远程复制功能，而且很难协调不同存储间的数据一致性关系。当不同阵列都归在 IBM StorWize V7000 下时，一切又恢复到比较简单、类似单台存储做灾备的环境。

传统模式下，一个数据中心起步阶段采用低端小存储，随着业务量增加，不断更新到更高端的存储上。这样不仅投入较大，而且每次升级对应用系统会带来一定风险及停顿（如数据从低端迁移到高端）。然而，IBM StorWize V7000 可以从低端起步，通过横向扩容（集群）的方式，增加控制器及容量，其可随数据及业务量的增长，平滑有序地升级成更高端存储系统。另外，IBM StorWize V7000 的外置虚拟化能力也带来极大升级空间，最大 32PB 的虚拟化空间足以满足大部分云存储的需求。

6.4.2 风险管理利器 2：EMC VNX 系列

从数量上来看，大数据的“可怕”之处首先就在于它的“大”，也就是数据的规模化效应，以现有的手动和人工的方式自然是不能够很好应对的，因此，重要的是要有高度自动化的解决方案来应对。

笔者注意到，市场上很多的产品都开始在简化管理界面、加强自动化与智能策略管理上下工夫，无论是如今正当主流的 IBM StorWize V7000 还是 EMC 推出的 VNX 系列，自动化程度都非常高。

EMC VNX 系列有两个分系列，分别是 VNXe 系列和 VNX 系列，VNXe 系列适用对象是中小型企业，VNX 系列的使用对象是大中型企业，如图 6-13 所示。因定位的不同，它们在所支持的协议、可扩展的接口、存储处理器 CPU 和内存（及缓存）、最大硬盘数和对复制软件的支持上都会有所不同。



图 6-13 EMC VNX 系列产品

EMC 最新发布的产品 VNXe 是一款整合程度更高的系统，它采用了新版本的 VNOX 操作系统，配备了一款双核英特尔处理器和 4GB RAM；在设置上更加简单，同时增加了 CLARiiON 和 Celerra 源技术所不具备的各项管理和支持功能。

专家提醒

以往，人们认识的数据修复技术往往是"回存"技术，就是要把备份数据介质倒回生产系统中，然后等待恢复的效果和业务的启动，这种技术存在众多风险。首先是在漫长的数据恢复之前，完全无法预料恢复时间和恢复可靠性。其次，一旦恢复成功，却发现恢复的数据并非自己需要的时间点数据，或者需要的数据不存在，这时已完全无法回退到初始状态，系统将进入更为严重的不可控状态。

VNXe 的易用性很强，配备了 Unisphere 向导设置程序、针对应用程序优化的管理功能以及 EMC 所说的一键帮助和支持功能，即用户只需一步操作即可进入自动诊断、服务状态及进入自助式用户社区。VNXe 产品以非常直观的管理界面，让用户可以通过七八步，在 2 分钟内为 500 个 Exchange 邮箱或 1TB 的 Vmware 数据存储配置好存储容量。

其中，VNXe 3100 采用 2U 或 3U 标准工业设计的机架式机箱，标配系统中除了附带用于 SAS 和 iSCSI 连接的 1Gbps 的以太网连接，还有 FlexI/O 插槽，其可提供额外的 1Gbps 端口，为扩展连接更多的设备并提高性能提供了先决条件。并且在容量方面，还提供简单的容量扩展，最大可添加 96 个 SAS 驱动器，按 1TB 的 SAS 驱动器容量计算，其最大可扩至 96TB 的存储容量。

自动化、块数据与文件数据的统一存储及虚拟化带来的存储系统整合，这些方法都能够有效降低数据存储尤其是大数据存储的风险。

6.4.3 风险管理利器 3：戴尔 EqualLogic 平台

如今，数据信息成为了商业价值的核心部分。由于实时获取数据、企业移动计算和虚拟化普及等需求的推动，预计从现在一直到 2020 年，企业存储每年将以 60%以上的速度增长，这一数字并不令人感到意外。平均来讲，每 18 个月企业数据便会翻一番。但实际上，似乎多数企业对管理数据增长做得不够好，而要依靠不见增长的预算来完成这一任务，这其中就存在很大的风险。

因此，用户可以考虑采用戴尔 EqualLogic 平台，其无缝扩展的架构和智能阵列软件，可以与企业第一层应用和虚拟环境自然集成，从而帮助企业高效地管理数据，却不会增加复杂性。EqualLogic 的自动化功能可以帮助企业每年将常见存储任务的管理时间大幅降低，将虚拟机（VM）部署提速超过 70%以上。

例如，EqualLogic FS7500 是唯一针对中小规模部署进行过优化的横向扩展统一存

储体系结构，借助它可以无中断地增大块和文件的容量，如图 6-14 所示。



图 6-14　EqualLogic FS7500

专 家 提 醒

大数据灾备系统的有效性问题涉及灾备建设的实际目标和符合目标的灾备技术路线，要清楚认识灾备系统的有效性问题，人们必须领悟到一个更深层次的道理：灾备系统的建设要求灾难防御全方位，不能只防小概率的自然灾害，更要防止概率大的设备故障和逻辑故障，严密的多方位防护网才是取胜之道。

未来与存储密切相关的两个挑战：一是非结构化数据的迅猛增长对于全球的企业用户而言都是一个相当头疼的问题；二是企业数据中心面临着向虚拟化、云计算转型的需求。毫无疑问，戴尔 EqualLogic 作为戴尔最重要的存储平台，必须要能完美地帮助企业迎接这些挑战，这样才能赢得自身的胜利。

6.4.4　风险管理利器 4：NetApp FAS 平台

NetApp FAS 系列产品的控制器承担了所有工作，包括 RAID、文件系统、网络 IO、双机集群（HA）系统等，它是一个完整的、一体的产品，如图 6-15 所示。





图 6-15　NetApp FAS 系列产品

下面以 NetApp 入门级 FAS2000 系列中的最新成员 NetApp FAS2240 为例，介绍 NetApp FAS 系列产品的主要特点，如表 6-1 所示。



表 6-1　NetApp FAS 系列产品的主要特点

主 要 特 点	细 节 说 明
性能和可扩展性	NetApp FAS2240 的性能比以往产品提升了两到三倍，因而灵活性也得以提高，便于客户最大限度地利用存储资源，支持要求苛刻的工作负载，并根据业务需求的变化添加增强功能
精简性	该管理工具简单、易于使用并随附于购买的系统中，其可帮助用户提高存储和服务效率以及生产率，并减少存储管理对有限 IT 资源的影响
NetApp Data ONTAP	FAS2000 系列运行最新版本的 Data ONTAP 操作系统，可为用户提供一个支持多种工作负载且具备高灵活性的可扩展统一平台，帮助他们满足不断增长的业务需求
可扩展的统一架构	NetApp 提供真正统一且可扩展的架构，支持客户轻松且经济地升级到更高端的系统和新功能，而无需执行“叉车式”升级。NetApp 的创新型统一平台可帮助用户构建高效灵活的可扩展基础架构，满足目前和未来的需求
行业领先的效率	NetApp 可提供行业领先的效率，因此中型企业的用户可从中受益。其他存储供应商只提供一两种存储效率技术，而 NetApp 提供 9 种集成的技术，可以帮助用户节省大量资金

6.5　大数据风险管理应用案例

大数据时代的来临，对中国来说面临安全管理能力、存储及处理能力、应用能力和人才培养能力等多方面的新挑战。对于很多企业来说，大数据并不意味着机遇或是商业上的无限潜力，在他们能够很好地管理数据之前，大数据只意味着风险和无穷无尽的烦恼。那么，如何解决大数据的风险和烦恼呢？本节主要介绍大数据风险管理的应用案例，希望对读者有一定的启发和学习价值。

6.5.1　【案例】“闪电计划”为数据护航

不久前，EMC 发布了传说已久的“闪电计划”，并推出了 VFCache，其旨在通过利用闪存的快速读写优势来加速数据流通速度，加强服务器与外部存储系统之间的联系，如图 6-16 所示。特别是针对关键应用环境中具有涡轮增压性能的服务器闪存缓存解决方案，通过提供线内重复数据消除功能，设立了企业闪存效率的新标杆。

同时，EMC 通过实现 VFCache 与 VMware® vSphere® vMotion 之间的新的互操作性，使得虚拟机可在由 VFCache 加快的环境中实现无缝、灵活地移动，这扩展了其在 VMware 环境的领导地位。EMC 继续投资于业界最全面的闪存产品组合，在保持网络存

储的高可用性、灾难恢复、数据完整性和可靠性等优点的同时，提供闪存具备的所有性能优势。



图 6-16　VFCache

在存储界中，磁盘阵列中采用 Flash 技术的磁盘通常被称为 SSD，随着对高性能的要求和 Flash 技术的价位的快速拉低引发了“caching tier（缓冲层）”。缓冲层是一个使用 Flash 技术的大容量第二级 cache，它位于服务器与存储磁盘之间。

EMC 的 VFCache 是一个面向服务器的 Flash-cache 解决方案，它运用了智能 cache 软件和 PCIe（Peripheral Component Interface Express，总线和接口标准）Flash 技术，旨在解决延时问题和加速带宽，最终可以极大地提高应用性能。VFCache 的技术亮点如表 6-2 所示。

表 6-2　VFCache 的技术亮点

技术亮点	具体说明
效率更高	EMC 正在发挥其在备份环境中的重复数据消除的领导能力，并将该技术应用到高速闪存缓存领域。通过更大的高效闪存对缓存数据进行线内重复数据消除，在“重复消除”收益很高的应用环境中，VFCache 的闪存缓存容量显著提高，并极大地延长了闪存卡的预期寿命
深度集成	在虚拟、存储和服务器层面上，VFCache 实现了更深度的集成，使关键任务应用环境最大化
涡轮增压的性能	VFCache 是当今最快的 PCIe 服务器闪存缓存解决方案。VFCache 被置于服务器中，热数据无需从网络穿过以到达存储阵列，这使吞吐量在某些情况下达到 3 倍的提升，并减少 60%的延迟。通过 PCIe 闪存卡实现更高的吞吐量和反应速度，需要的 CPU 和内存资源却比竞争产品少 4 倍
操作环境自动化	VFCache 与 VMware vSphere vMotion 之间的互操作性，使其更快、更易于实现持续正常、流畅地运行，以及完整的环境维护，并使迁移顺利进行，这有助于客户加快其云计算之旅
智能缓存策略	VFCache 在服务器上实现了新一层的高性能存储。VFCache 将 EMC FAST 架构延展到支持一个智能的端到端的数据分层和存储到服务器的缓存策略

续表

技 术 亮 点	具 体 说 明
性能更佳	VFCache 的最新版本支持每个服务器有多块 PCIe 卡，并提供更多容量选择，可支持新的 700GB PCIe 卡以缓存更大的工作集，并为客户提供更优性能，可通过调整 VFCache 缓存算法进而降低延迟时间
企业级数据保护	VFCache 通过将全盘数据“透写式缓存”到存储阵列使客户受益，使数据拥有可用性、完整性、可靠性和灾难恢复的存储解决方案。无需任何冗赘的存储，这些信息依然可分享和可扩展

【案例解析】: 在本案例中，VFCache 的发布使 EMC 成为第一家运用 PCIe 闪存技术帮助客户以合理的成本，满足客户需要的数据保护和数据智能，来确保其关键应用达到新的性能高度的公司，为大数据项目风险管理构筑了一道坚实的“城墙”。

当前，我国大数据存储、分析和处理的能力还很薄弱，与大数据相关的技术和工具的运用也相当不成熟，大部分企业仍处于 IT 产业链的低端。我国在数据库、数据仓库、数据挖掘以及云计算等领域的技术，普遍落后于国外先进水平。笔者认为，我国如何借用国外先进的技术平台，借用其对大数据资源的存储和整合能力，实现从大数据中发现、挖掘出有价值的信息和知识，是当前我国大数据存储和处理所面临的挑战。

6.5.2 【案例】智慧存储化解大数据风险

服务器与存储融合的趋势日趋明显，而纯粹的存储厂商做服务器闪存卡更是有代表性的大事件，EMC VFCache 一道“闪电”拉开了存储大佬们的闪存之争的序幕。虽然 IBM 已有 eXFlash 这样的闪存技术，但是在这场争夺战中，IBM 似乎显得有些低调。

那么，对于 IBM 这样既有服务器又有存储业务的厂商来说，在大数据方面又有怎样的动作呢？为了帮助更多企业把握“大数据”机遇，化解大数据在企业内部的风险叠加，IBM“智慧存储”战略帮助企业 CIO 更加有效地收集并提取信息，合理分析并加以利用，借助这种更加灵活、高效和简单的方法管理企业信息架构。

例如，IBM 近期提高了多个产品的效率和性能，如表 6-3 所示。

表 6-3　IBM 近期增强的产品和策略

产 品 策 略	增 强 方 面
面向中小企业的 IBM System Storage DS3500 及采购高密度设计、可构建高性能计算环境的 DCS3700	这些产品现已具备增强型闪速复制功能，能够多复制 50%的快照，从而加快备份速度；此外，精简调配功能可将未用容量保存在存储资源池中，以便按需提供给应用使用，从而能够提高磁盘存储器的利用率，同时降低存储成本

续表

产品策略	增强方面
IBM 磁带系统库管理器 IBM Tape System Library Manager（TSLM）	能够给客户提供多个磁带库的单一综合视图，从而扩展 IBM TS3500 磁带库的使用范围并且简化其使用流程。TSLM 能够与多代企业级和 LTO 驱动器及介质互操作，从而将数据保存在单一磁带储备库中，并且允许企业通过 IBM Tivoli Storage Manager 集中管理这个磁带库
IBM 线性磁带文件系统（LTFS）存储管理器	允许客户使用 IBM LTO 5 磁带库及 IBM LTFS Library Edition 针对大型视频文件等多媒体文件实施生命周期管理，从而显著降低视频档案的许可成本及录像带介质成本
IBM Tivoli Storage Productivity Center（TPC）套件	TPC 的全新增强特性将允许公司更好地满足大数据存储需求。通过基于 Web 的全新用户界面，TPC 能够从根本上改变 IT 经理查看和管理存储基础架构的方式。此外，将 TPC 与提供直观报告与建模功能的 IBM Cognos 相集成将允许客户轻松创建高质量的特殊报告和定制报告，以便做出更加明智的决策。TPC 采用简单包装方式，允许客户通过单一许可开展全面的管理、发现、配置、性能保证和复制工作
智慧存储方法	进一步改进智慧存储方法，将 IBM Easy Tier 功能扩展到基于服务器的直接连接 SSD 领域，以便帮助客户协调磁盘系统与服务器之间的数据迁移活动，如图 6-17 所示。IBM Easy Tier 可基于策略和活动将数据自动转移到最适合的存储位置，包括多层磁盘和 SSD

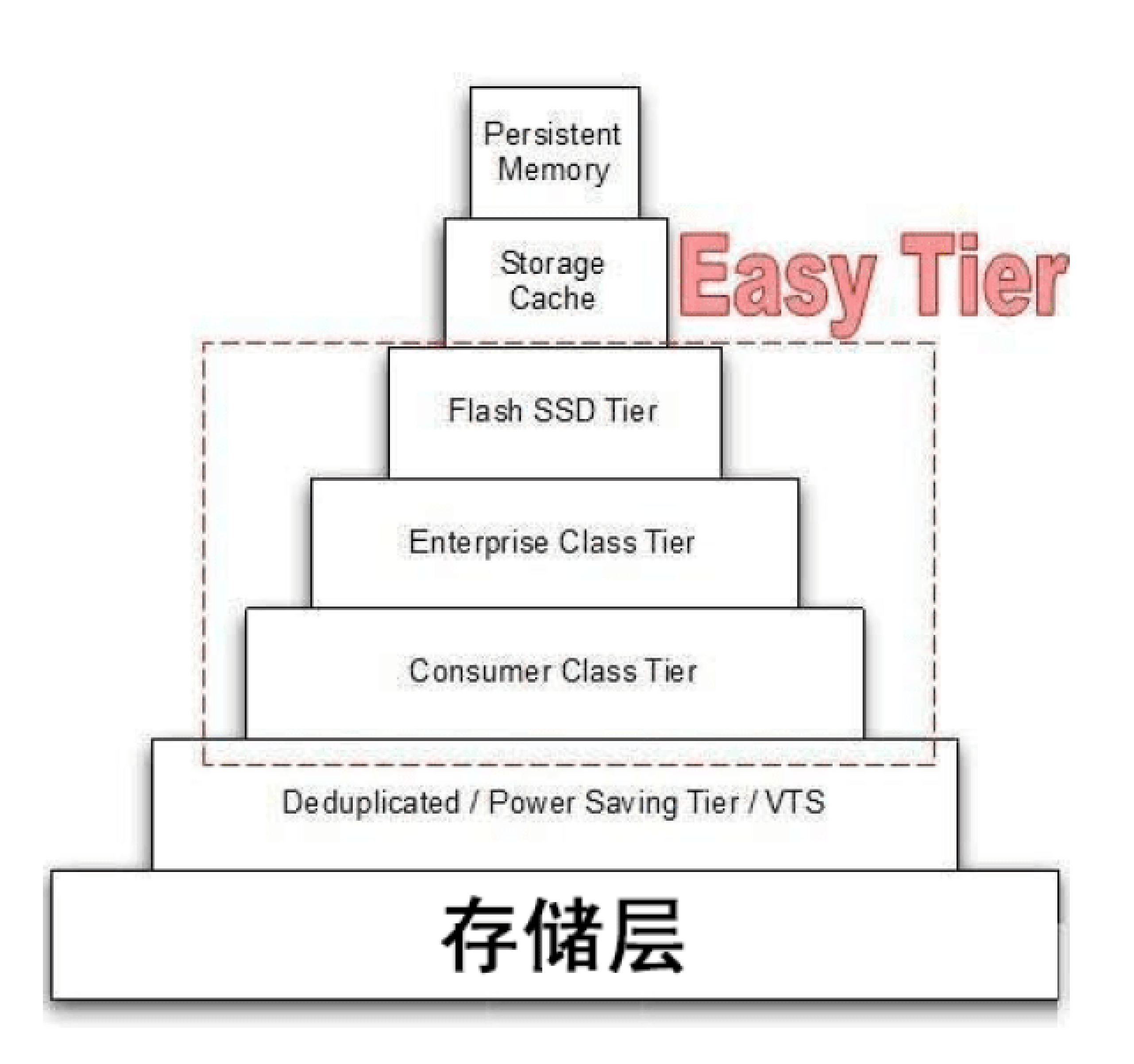


图 6-17　IBM Easy Tier 可支持 3 个存储层

【案例解析】：在本案例中，IBM 作为领先的 IT 服务提供商，已经紧紧抓住了发展趋势，利用自身优势、资源及解决方案深入企业业务需求，帮助企业认清方向，通过“智慧存储”战略解除企业数据危机并实现新时期的智慧成长。

与 IBM 的主要业务相比，我国在大数据存储和分析方面都存在缺陷：

➢ 在大数据存储方面，数据的爆炸式增长，数据来源的极其丰富和数据类型的多种多样，使数据存储量更庞大，对数据展现的要求更高。然而，目前我国传统的数据库，还难以存储如此巨大的数据量。

➢ 在大数据的分析处理方面，由于针对具体的应用类型，需要采用不同的处理方式，因此必须通过建立高级大数据的分析模型，来实现快速抽取大数据的核心数据，高效分析这些核心数据并从中发现价值，而这些数据分析能力我国还很欠缺。

因此，笔者建议那些经过激烈市场洗礼的企业在全新 IT 环境下更要抓住机遇，做出明智决策，大数据带来的全新 IT 挑战将成为企业基础架构变革的动力。

6.5.3 【案例】谷歌循环利用“数据废气”

拼写检查对于英语写作来说是很重要的一个纠错功能，Google Docs 的文档已经支持拼写检查，而且现在使用 Google Docs 的表格也可以接受拼写检查了。如图 6-18 所示，左侧是新的系统，右侧是老的系统，Google 终于意识到自己的“Gmail”也是一个正确的拼写单词了，因为新的系统结合了 Google 的在线拼写检查功能，而老系统只是比照词典去查错，字典里显然没有 Gmail 这个词。

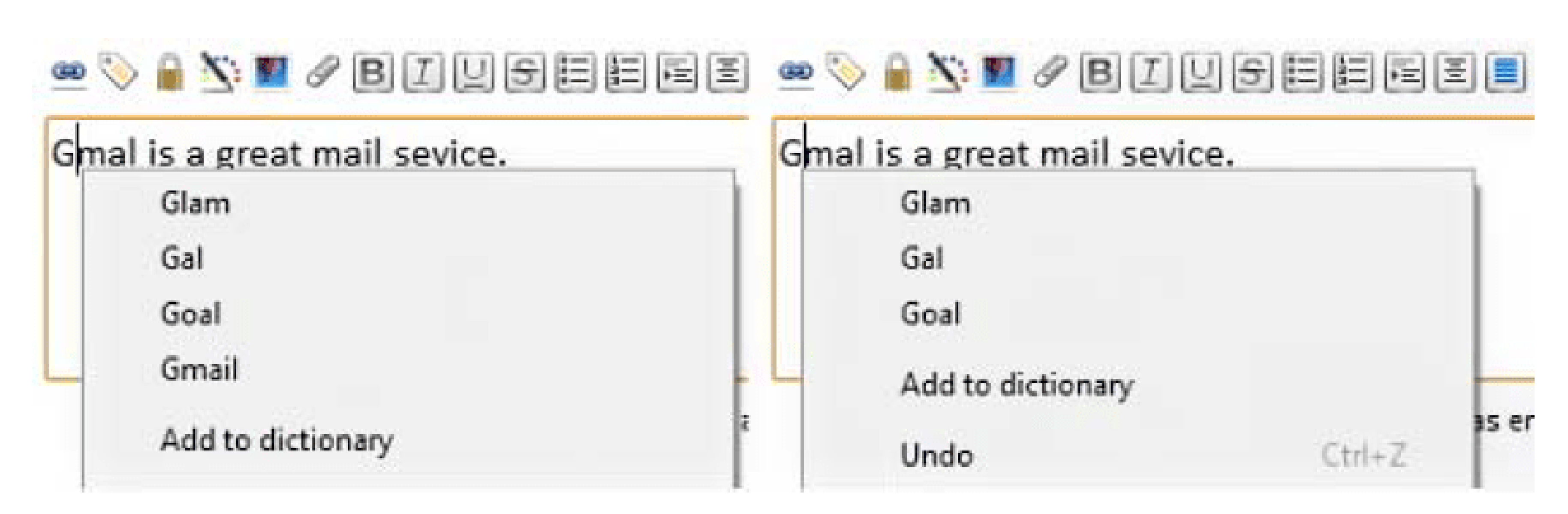


图 6-18 Google Docs 新老系统对比

由于人类的语言极其复杂而且内容繁多，有非常多的规则需要设计，因此造成同一句话可以表达不同意思，不同的话可以表达相同意思，以及流行语更新很快等问题。因此，一直以来，专业的拼写检查器（spell checker）很难达到人们的应用要求，比较起来，搜索引擎成为了最先进的拼写检查工具。

很多人都有过这样的经历：对于一个句子、单词、成语甚至古诗不确定的时候，就拿 Google 或者百度搜一下。有意思的是，不管 Google 还是百度都不是作为拼写检查器被设计出来的，而且他们也没有专门的“拼写检查”功能。之所以这个歪打正着的功能

居然这么好用，是因为它收集而且组织了极其大量的信息。

在大数据时代，搜索引擎能看到所有人们提出的问题，所以如果你在拼写中或者用词中犯了一个错误，它能通过比对海量数据来预测出你的这个错误，从而导致搜索引擎成为了目前为止最先进的拼写检查器。

这些用户之间交互的语言“碎屑”却被谷歌当成了金粉，收集在一起就能锻造成一块闪亮的金元宝。一个用来描述人们在网上留下的数字轨迹的艺术词汇出现了，这就是“数据废气”，它是用户在线交互的副产品，包括浏览了哪些页面、停留了多久、鼠标光标停留的位置、输入了什么信息等。许多公司因此对系统进行了设计，使自己能够得到数据废气并循环利用，以改善现有的服务或开发新服务。

【案例解析】：“数据废气”向来被人们当成是一种负担，累积在一起将会带来极大的存储压力。但从本案例继续往下分析，可以看到“数据废气”将成为公司的巨大竞争优势，相同的方法和原理在人工智能、预测分析学的很多其他方面都有着应用，例如人脸识别技术等，这些应用的基础只有一个——那就是极其大量的数据。因此，把 Google 当拼写检查器使用，这个有趣的现象值得我们好好去观察和思考，也许海量数据真的会带来人工智能的新时代。

6.5.4 【案例】借助淘宝大数据控制风险

做服装生意的 90 后美女小丽最近开了一家淘宝店，但等了两个多月才等来第一单生意。小丽问第一个客人为什么没人来她的网店购物，顾客告诉她，小丽的网店页面上显示没有交保证金，所以买家觉得她的店不那么“靠谱”。

听说可以用保费代替保证金，小丽马上买了这款保险，“只花 30 块钱就能帮我提高信用，为什么不试一下？”

2013 年 11 月 25 日，众安保险联合阿里巴巴推出“众乐宝—保证金计划”（以下简称“众乐宝”），其将利用淘宝全量大数据进行风险控制。

众安保险的定位为数据公司，掌握的大量稀缺性数据是众安保险的价值之一，公司要做的是锤炼团队对数据的分析、运用能力。“众乐宝”将通过严格的事前风险控制，有效识别风险客户，并根据卖家的信用表现给予匹配其信用的承保额度，通过实时的控制监控跟踪卖家的风险，且在事后针对恶意风险客户给予信息披露。

“众乐宝”改变了淘宝卖家交“保证金”的惯例。此前，在淘宝开店，卖家需缴纳 1000 元～10000 元不等的消费者保障基金，卖家并不能动用这笔保障基金。不过“众乐宝”推出后将改变以往卖家交“保证金”的方式，卖家可以选择不交 1000 元的保障金，改为投保一年“众乐宝”，一年费率为 3%，如图 6-19 所示。一旦卖家发生违约行为，保险公司先行垫付赔偿买家，然后向卖家追偿。



图 6-19　众乐宝—保证金计划

“众乐宝”正式上线运营后，淘宝卖家只要缴纳 18 元就可以获得保障额度为 1000 元的半年期“众乐宝”保险。这样的理赔形式，对于卖家来说，提高了资金的使用效率；对于买家来说，保险的先行赔付可以缩短维权过程，这能更好地提升买家的购物体验。

很多卖家都知道，在淘宝的搜索排名中，有没有缴纳保证金也是影响因素之一。而对买家而言，这个店铺有没有消保标志，同样会影响他们的购买行为。淘宝网相关数字显示，至少有 500 万左右的淘宝卖家没有缴纳保证金，而“众乐宝”的首选目标客户无疑是这些没有参加消保的卖家。因为对于他们而言，一笔极低的保费就可以获得消保标志。同时，众安保险也极力争取已经缴纳保证金的卖家，对这些卖家而言，用少量的保费就可以盘活其被冻结的保证金。

在此过程中，“众乐宝”的风险主要来自两部分，即卖家本身的信用风险和卖家本身的经营风险。众安保险会在事前对卖家的信用以及经营情况等进行信用评估，并采集了淘宝的全量数据来对卖家做信用评估。

【案例解析】：在具体操作中，目前线下数据较为碎片化，线上数据则更为透明，可逐渐完善信用平台。在本案例中，众安保险不仅可以根据卖家的交易记录，还可根据买家对卖家的评价来测算卖家信用，大数据将全程应用于“众乐宝”风险控制的各个阶段。



读书笔记



精准行业聚焦篇

7

平台：信息通信大数据

学前提示

移动互联网发展起来后，数据爆发性增长，运营商怎样利用好手中的大数据？如何进一步优化、升级网络，以应对“大数据”时代用户的流量需求呢？大数据时代运营商面临的是机遇还是挑战？本章将结合传统通信行业，介绍大数据的解决方案和应用案例。

要点展示

- 信息通信平台大数据解决方案
- 信息通信平台大数据应用案例

7.1 信息通信平台大数据解决方案

车联网、物联网、云计算、移动互联网等以及遍布全球的各种各样的传感器，无一不是数据来源或者承载的方式。大数据的累积效应正给整个 IT 业带来变革。特别是云概念和 3G 的深入发展，各大运营商面临着越来越大的数据压力，同时 IDC（Internet Data Center，即互联网数据中心）扩容，偏向以存储为主的云服务。

对于运营商来说，这个“大数据”主要是大量的用户行为数据。随着智能手机的普及，运营商将获得更加完备的用户行为数据，而能否挖掘出这些数据的价值将决定运营商能否把握住大数据带来的机遇。

7.1.1 运营商在大数据时代的认识转变

移动互联网时代的到来带动了通信业新的变化，以腾讯、阿里巴巴、百度、奇虎 360 等为代表的互联网公司目前已经形成了与传统电信运营商价值链重新划分的格局，使得运营商的角色正在不知不觉中发生着变化。

不管用户换什么 OTT 平台和终端，数据总归会流经管道和运营商。所以有人问，淘宝也有大数据，腾讯也有大数据，运营商的大数据和他们有何区别呢？其实，区别在于，淘宝拿不到腾讯的大数据，腾讯拿不到淘宝的大数据，但运营商可以同时拿到淘宝和腾讯的数据，只要有这个必要。

专 家 提 醒

OTT 是 Over The Top 的缩写，是通信行业非常流行的一个词汇，这个词汇来源于篮球等体育运动，是“过顶传球”之意，指的是球类运动员在他们头上来回传球而使其到达目的地。OTT 在商业中的意思是，互联网公司越过运营商，发展基于开放互联网的各种视频及数据服务业务，强调服务与物理网络的无关性。互联网企业利用运营商的宽带网络发展自己的业务，如国外的谷歌、苹果、Skype、Netflix，以及国内的 QQ、阿里旺旺等。不少 OTT 服务商直接面向用户提供服务和计费，使运营商沦为单纯的“传输管道”，根本无法触及管道中传输的巨大价值。

当前，通信业务的竞争日趋激烈，保证网络质量无疑是网络运营商竞争取胜的关键所在。为提高网络服务质量，网络运营商必须建立高效运作的维护体系，推进移动网络基础运营的精确管理，并以信息化为支撑，通过先进的维护手段不断提高维护管理效率，为整个运营网络提供可靠的业务保障。那么，大数据的到来对运营商有什么启示呢？笔者认为至少有以下两点：

（1）业务类型的转变。传统运营商所提供的服务类型已经从单一的话音结合少量

的数据通信，向多媒体、IPTV 等多业务叠加模式演变。

（2）业务价值链的改变。在大数据时代，运营商不得不面对为数众多的并且在逐步壮大的互联网服务提供商和应用提供商，运营商想自己直接经营这些业务显然不太现实。因此，如何处理与互联网公司的关系？公司化运作、新的 IT 技术的利用是否是其转型的救命稻草？云、管、端三线布局能否解决管道化的忧虑？这是大数据时代摆在我国运营商面前的难题。

专家提醒

在需求不断变化增长的发展趋势下，很多运营商在尝试布局“云管端”架构，如图 7-1 所示。

- 云：云平台将成为未来信息服务架构的核心。
- 管：超宽带智能网络是实现该新架构的基础和前提，同时是实现“云-端”互动的桥梁。
- 端：融合终端（Terminal，集中式主机系统）的智能化，将大规模地在各行业得到应用。

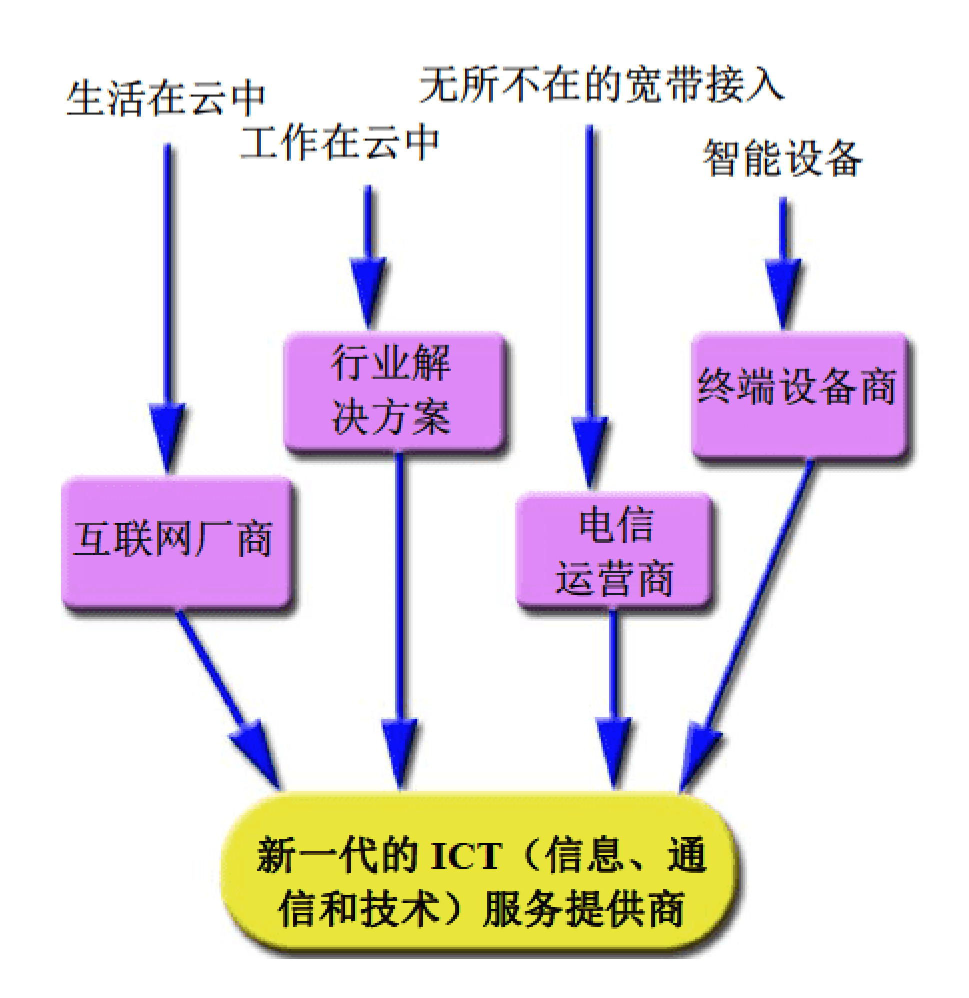


图 7-1　不断变化增长的通信市场需要新的“云-管-端”模式

7.1.2　运营商在大数据时代的模式转型

移动互联网发展起来之后，运营商在近两三年开始关注大数据。大数据不是新的概念，在移动互联网发展起来，数据增长速度加快，整个产业压力突出，传统数据库技术已无法满足运营商对大数据充分利用的需求的背景下，大数据成为近年来的热点。但是，对运营商来说，数据爆发性增长后，并没有为其带来可观的收入。

究其原因，主要有以下两点：

（1）运营模式受限。由于大数据产业具有强烈的互联网特征，因此运营商现有的

运营模式很难帮助实现大数据产业的迅速发展。

（2）组织结构过时。对于大数据产业，运营商传统的金字塔式的组织结构已经过时，传统架构的信息系统及组织架构已无法应对海量数据和创新型应用，那种由上而下的运营模式无法更接近用户的需求，显然已经阻碍运营商自身大数据产业的纵深发展。

尽管大数据在商用道路上的发展困难重重，但是由于运营商有经营大数据的先天优势，且又有在互联网时代沦为“数据管道”的压力，还有大数据时代信息价值的高昂，使得探索和发展大数据成为运营商最明智的选择和最好的出路。

总的来说，运营商运用大数据主要有 4 种模式，如表 7-1 所示。

表 7-1　运营商运用大数据的模式

运 用 层 面	具 体 操 作
市场	运营商可以利用大数据对自身的产品进行服务，通过大数据分析用户行为，改进产品设计，并通过用户偏好分析，及时、准确地进行业务推荐，强化客户关怀，这样就可以不断改善用户体验，增加用户的信息消费以及对运营商的粘度
网络	可以通过大数据分析网络的流量、流向变化趋势，及时调整资源配置，同时还可以分析网络日志，进行全网络优化，不断提升网络质量和网络利用率
企业经营	可以通过业务、资源、财务等各类数据的综合分析，快速准确地确定公司经营管理和市场竞争策略
业务创新	可以在确保用户隐私不被侵犯的前提下，对数据进行深度加工，对外提供信息服务，为企业创造新的价值

只要做到以上 4 种模式的转变，运营商即可借助大数据来实现从网络服务提供商向信息服务提供商的转变。笔者认为，运营商应该跳出互联网看互联网，将大数据作为重点业务发展领域，毕竟运营商拥有的“数据矿产”资源是任何其他企业所不具备的，运营商应该基于大数据的基础发展延伸业务。

专 家 提 醒

在大数据时代，运营商必须根据市场需求，全面转向以客户和消费者为中心的运营体系，重新梳理企业的经营模式和组织架构，这就是模式的创新。

7.1.3　运营商在大数据时代的机遇前景

运营商手中的“大数据”如同一座丰富的金矿，然而对其价值的挖掘却由于体量太大的缘故迟迟无法有效推广，如图 7-2 所示。

1. 运营商为何难以下手

当谈到大数据话题时，通信运营商们都不愿公开谈论他们的进展。这表明运营商或者是在部署独特的亦或是商业敏感性的解决方案，又或者他们还未下定投身大数据的决

心。笔者认为，在营运商的大数据道路面前，至少有以下两道坎：



图 7-2　运营商如何挖掘“大数据金矿”

（1）市场没有定型。由于国内还没有成熟的市场，所以国内运营商在大数据的商业挖掘上还没有看到应用的出现。通常大家能看到的一些与位置有关的服务，例如餐饮、活动查询等，其实与运营商的关系并不大，一般是通过 GPS 定位来实现的。

（2）政策监管是空白。运营商所掌握的用户信息是十分精确的数据，不仅仅是用户的身份信息、手机号码等，甚至连用户的所处位置、通话状态等都能够获取。在通信行业里，通话记录等属于涉密信息，在这个信息的获取上是没有灰色地带的，如果没有政策导向，一味只考虑利用用户信息挖掘商业价值，就会面临信任危机。

2. 从云计算来打“首战”

运营商在云计算和大数据应用的发展上，相比较互联网企业有一定的优势，利用好了，找准了发力点和突破点，在移动互联网产业的发展中可占据一席之地。运营商发展云计算的先天优势是其在电信时代所积累的遍布全球的 IDC（数据中心）和庞大而详细的用户数据（包括身份数据和行为数据），而且都是电信级的质量和品质。运营商的 IDC 不仅可以满足自身业务的需求，也可以为互联网企业提供相关租赁、托管等服务。

运营商 IDC 众多，对带宽绝对控制，有国有资产的公信力，无论发展公有云、私有云还是专属云，均具备优势。在云计算的发展中，平台才是王道，“得平台者，得云计算半壁”。

运营商应与开发者合作共赢，从以自己单独运营为主逐渐转向专注提供开放的、低门槛的开发平台和环境，汇聚广大开发者共同开发。当然，运营商发展云计算，不能仅停留在云计算本身上，也不能仅停留在云计算基础设施建设上，而是要专注于云计算应用，使其落地开花。

因此，运营商可以利用自身优势，有针对性地搜集各种不同类型的数据，打好时间差，先发制人，可以获得先发优势。否则，随着人们的行为越来越多地发生在互联网公

司端，互联网公司搜集到的数据越来越全面，运营商的优势将不复存在。另外，运营商要学会降低成本，保证合理的质量，并进行市场普遍定价，这是运营商必须考虑和解决的问题。

专家提醒

运营商的自身优势主要有以下几点：

- 可以看到用户的年龄、品牌、资费、入网渠道，还能够看到他们的上网时间、上网地点、浏览内容偏好、各种应用的使用时间等。
- 能够知道用户用了什么样的终端，包括 IMEI、MAC、终端品牌、终端类型、终端预装了哪些应用、终端的操作系统、终端的尺寸等。
- Web 浏览记录、传感器信号、GPS 跟踪和社交网络信息等数据也都会被运营商掌握。

从这些数据中分析用户的行为习惯和消费喜好，正是大数据的精髓所在。

3. 逐步进入大数据领域

过去，运营商已经积累了大量的优质数据，但其价值一直未被发现。如今，大数据时代的到来，使这些数据反倒可以成为运营商“咸鱼翻身”的利器。目前运营商的优势只是数据大，需要将数据大变成大数据，对数据进行充分的挖掘和分析，并从中生发出新的业务形态和价值来。

（1）扩大现有的数据业务。运营商要接受大数据带来的变革性影响，顺应数据业务主营化的大趋势，将数据业务及时转换成自己的主营业务。电信业原有的主营业务是语音业务，数据业务只是辅助性业务。但在移动互联网中，数据业务上升为主营业务，有的甚至可以占到 76%以上，而语音业务成为副业。

（2）初步构建大数据系统。大数据时代，运营商可以提供用于云服务的数据融合技术、海量数据挖掘技术和大规模分布式技术。围绕新核心系统 BDS 这个中心，形成运营商的网络大脑，进而建立网络数据子系统、用户数据子系统和业务数据子系统。且其 IDC 有天然优势，不用求人。这一部分，从互联网角度看，也属于运营商最优质的资产，可以成为移动互联网数据核心业务的重要组成部分，甚至是重心所在。

（3）认清大数据发展方向。运营商将来努力方向是完善面向客户的支撑系统，全面提升面向客户的支撑能力。不应局限于传统 IDC 思路，只把重点放在服务器托管、出租设备等方式上，还需要深入到业务内部，思路向数据方向转变，提高服务的能力。

（4）应用才是真正的财源。移动互联前沿的竞争在于除了提供 IT 服务之外，还要与应用结合起来，提供基于应用的云计算服务。例如数据采集之后，要把数据业务展开成几个具体的产业；再如数据增值前，可以增加咨询加工服务，再往下是平台业务、很多分散的应用，这恐怕不是运营商一家能够做得到的，可通过合作做大产业。

专 家 提 醒

应用在面向对象上，通常可分为个人用户应用（面向个人消费者）与企业级应用（面向企业），在移动端系统分类上主要包括 iOS App（如同步推等）和 Android Apk（如 AirDroid、百度应用等）。

7.1.4　运营商在大数据时代的应对方案

运营商拥有丰富的大数据资源，包括数据资源、基础资源和平台资源，这些资源优势是其他企业无法比拟的。不过，这些数据只有经过长期的运营、使用和剖析，才能够真正发挥价值，如图 7-3 所示。

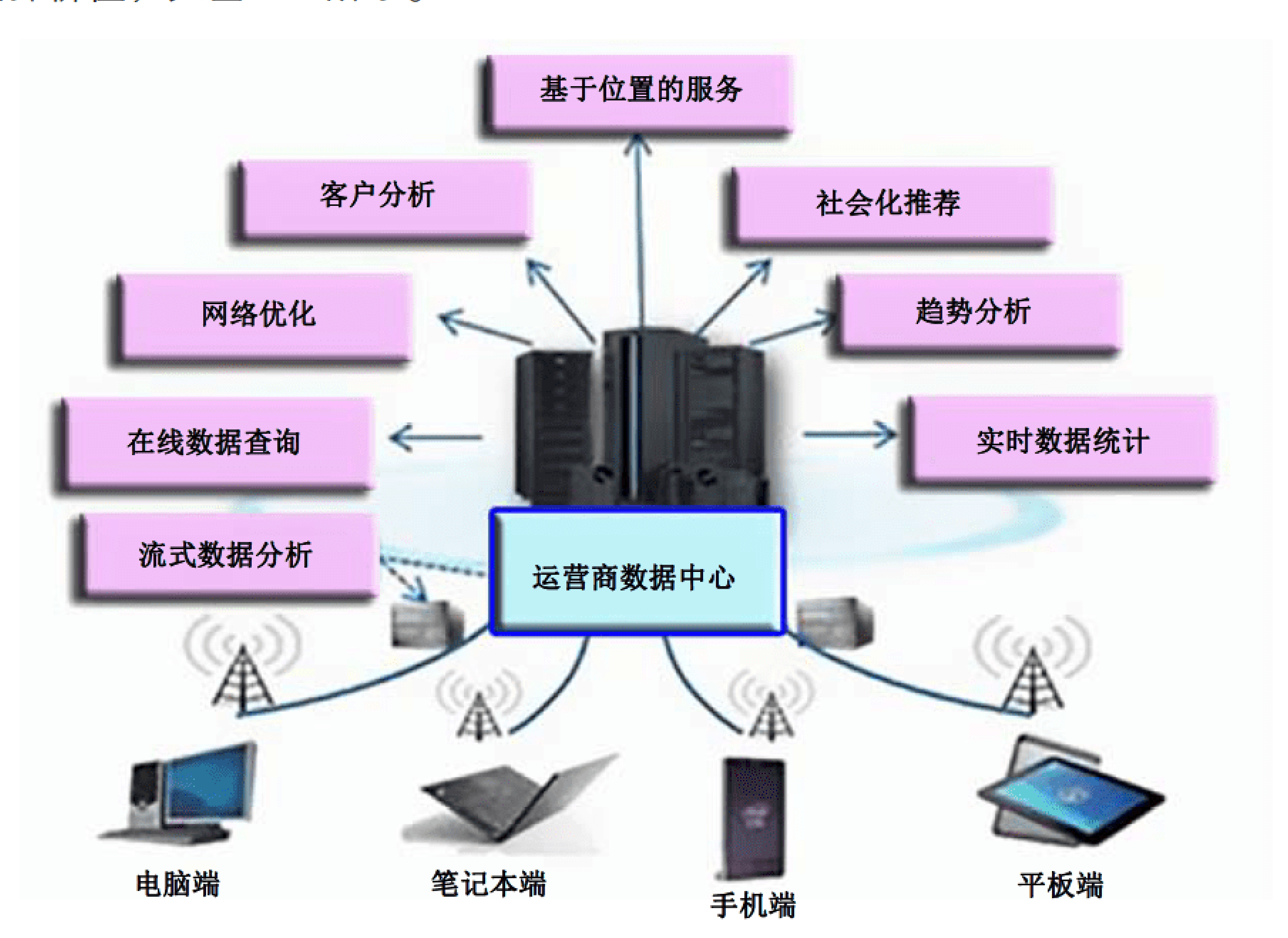


图 7-3　运营商大数据解决方案

（1）打造实时营销解决方案。运营商应整合现有数据建立数据集市，利用实时处理大数据的能力，打造基于数据的实时营销解决方案，提升企业销售服务能力。大数据处理分析平台的优势在于对海量数据处理的实时性，技术优势可以有效地保障实时营销解决方案的实施。例如，“基于位置的服务”是根据用户位置轨迹信息推送自有业务或者合作商家的产品信息，如对接近某大型商场的用户推送商店优惠信息，吸引客户消费。

（2）成为数据信息的融合者。运营商可以利用自有的品牌优势打造权威指数类产品，为客户的决策提供参考依据，可以提供更加全面、详尽、客观的产品，对于分析中欠缺的数据可以同其他行业进行合作共同挖掘数据中隐含的价值。

（3）提升其他行业的数据价值。电信运营商可为智慧医疗、智能交通、智慧物流、智能制造等领域提供解决方案，提升数据价值。

专家提醒

例如，交通管理行业在大数据时代，需要解决基于大数据及时查询、及时分析等业务需求。电信运营商可以利用如"全球眼"等业务和云存储方面的技术积累，提供海量交通数据的存储、分析、应用，同时利用智能管道进行交通信息的及时推送，这样可以更加有效地保障交通管理行业的及时性要求。

7.2 信息通信平台大数据应用案例

大数据并非运营商独家的概念，它已成为整个互联网行业共同关注的领域。互联网服务对传统通信运营商业务构成的冲击，反而可以加速运营商的转型，并催生新的机遇和市场空间。大数据恰恰就是在这种产业变化的情况下催生出的新业务，对于运营商来说，在大数据领域可拥有比传统基础电信业务更大的市场空间。本节主要介绍信息通信平台大数据的应用案例，希望对读者有一定的启发和学习价值。

7.2.1 【案例】西班牙电话公司的数据再利用

2012年10月9日，西班牙电信成立了名为"动态洞察"的大数据业务部门Telefonica Dynamic Insights，希望借此把握大数据时代商机，创造新的商业价值。

西班牙电信此次成立的大数据业务部门隶属于该公司此前成立的数字业务部门Telefonica Digital。大数据部门面向全球运营，主要目标客户为企业和公共事业部门，其将为客户提供信息和分析打包业务，帮助客户把握重大的变化趋势。

大数据业务部推出的首款产品智慧足迹（Smart Steps，如图7-4所示）就是将匿名的移动网络数据提供给零售企业等客户，让其了解在某个时段、某个地点的人流量，据此决策新店的选址、进行时段促销等。

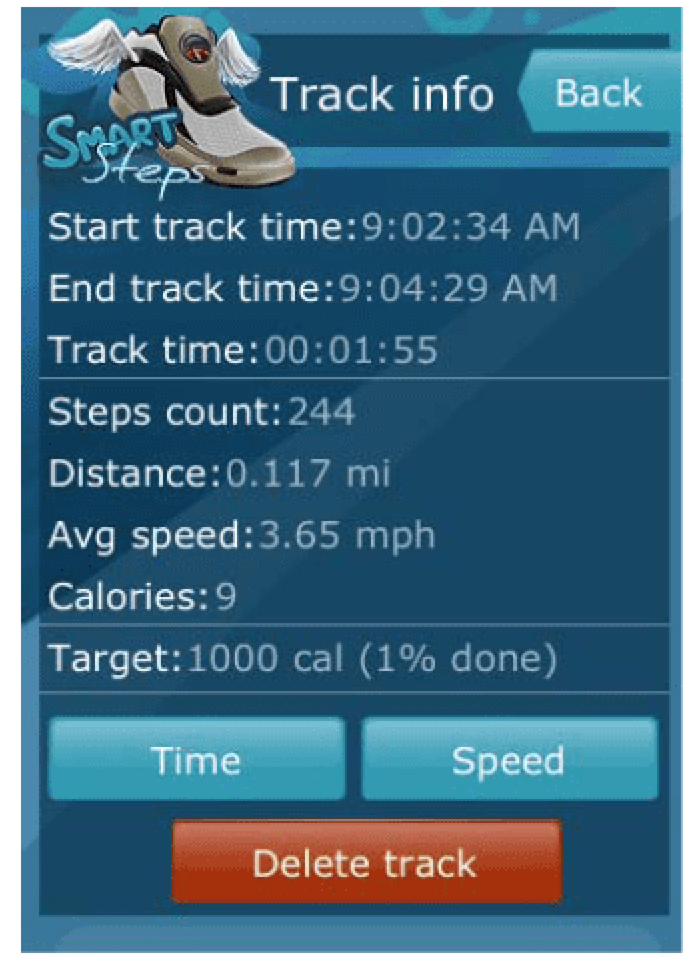


图7-4 Smart Steps界面

其实，西班牙电信在数据能力商业化领域已经进行了不少探索。例如，2011年1月，西班牙电信旗下英国O2运营商就在英国推出了免费WiFi服务，尝试将收集来的用户数据用在媒体广告和营销服务方面。免费的WiFi服务意味着更多的人会使用

这个服务，进而 O2 运营商就会收集到更多的用户数据，而广告商就能够利用这些数据进行更精准的广告投递。

2012 年，西班牙电信公司与 GFK 市场研究公司联手，成立新部门——西班牙电信数字洞察（Telefonica Digital Insights），以此获得德国、英国和巴西等市场的相关数据。

【案例解析】：大数据是数字经济建模的关键之一，是转换企业和社会每一部分又智能又可靠的方式，有促进经济增长、改善人们生活水平的潜力。在本案例中，西班牙电信通过 APP 应用对手机用户的一般活动进行定位，这不但有助于零售商作出战略决策，还可以帮助市政府制定停车场计划、管理公共事务。

笔者认为，大数据是对技术的综合应用，运营商要有开放、融合、服务和创新的心态，在大数据领域创造另一片天地。例如，一个大数据的应用通过收集数据，对大量图片进行分析，最终形成一个场景图。这就是对数据分析、统计技术、图片处理技术和人工智能合成技术的综合运用。

7.2.2 【案例】德国电信的大数据营销新策略

德国电信 T-Systems 是 SAP 第一批合作商，现已成为 SAP 认证的 SAPHANA 企业云运维服务供应商。T-Systems 作为德国电信子公司，通过对特定的 SAPHANA 平台基础设施的建设，已可提供基于云计算的端到端大数据服务。

T-Systems 的信息通信技术部主任 Olaf Heyden 说，"大公司对云计算越来越感兴趣，高效数据中心的需求在几年之后会越来越明显。"

此前，T-Systems 公司与英特尔公司在慕尼黑共建了试运行数据中心。两家公司对运行服务环境的可持续性和高效性进行了研究。正是基于这份研究结果，T-Systems 公司决定新建云计算数据中心。

德国电信 T-Systems 凭借在 SAPHANA 领域的专业知识，为客户提供大数据环境下高性能商业智能应用程序。企业通过该程序进行实时海量数据分析，并将结果作为"智囊"以供管理层参考。通过使用 SAPHANA 企业云，企业无需购买德国电信 T-Systems 相关"端到端"大数据解决方案和技术设施，只需使用建立在多样化云平台（DCP）上的应用程序便可轻松享受大数据的核心价值。

SAPHANA 平台除了可以快速处理大数据外，还支持全新的一体化分析方式，分析结果能够直接作为业务决策的参考甚至产生新业务，使得企业能更容易地满足阶段性需求。

专家提醒

SAP 提供一系列前所未有的新型企业应用，其中结合了大量交易与实时分析能力，能够显著优化现有的计划流程、预测流程、定价优化流程等数据密集型流程。HANA 是一个软硬

件结合体，可提供高性能的数据查询功能，用户可以直接对大量实时业务数据进行查询和分析，而不需要对业务数据进行建模、聚合等。

【案例解析】：SAPHANA 平台提高了对结构性大数据分析的能力。在数据中心、网络、应用程序和流程集成的完美配合下，SAPHANA 能够发挥全部潜能。在本案例中，德国电信 T-Systems 对于 SAPHANA 的性能进行了精准的投入，同时也已完成 SAPHANA 与多种基于云的 SAP 解决方案的一体化，这意味着相关的业务流程可以获得全面的改进。

聪明的决策来自于分析新的数据源，并用其增强现有的利用操作型系统和数据仓库中的结构化数据建立的分析和预测模型。大数据产品强调对传感器数据、网页日志数据、SNS 数据、文档等多种非结构化数据的分析。运营商可以将自己的业务技能和技术技能组织在一起，深入分析大数据，找到改善当前业务分析和预测分析的模型，并发现新的商业机会。

7.2.3 【案例】Verizon 利用大数据精准营销

威瑞森电信（Verizon）是美国最大的无线通信提供商和本地电话交换公司，该公司也是全世界最大的黄页印刷公司和在线黄页信息提供商，在美国、欧洲、亚洲、太平洋等全球 45 个国家经营电信及无线业务。

2012 年 10 月初，Verizon 成立了精准营销部门 Precision Marketing Division。根据部门副总裁 Colson Hillier 的介绍，该部门提供以下 3 方面的服务：

（1）精准营销洞察（Precision Market Insights）。提供商业数据分析服务。该服务已经开始向第三方售卖 Verizon 手上的用户数据，对商场、体育馆、广告牌业主等出售特定场所手机用户的活动和背景信息。

专 家 提 醒

Precision Market Insights 的具体做法如下：

Verizon 收集包括位置和 Web 浏览信息在内的用户数据，并将这些信息发给数据库，与从第三方拿到的人口统计数据（年龄、性别等）结合起来，Precision Market Insights 服务将数据进行聚类，然后卖给体育场馆、商场等需要做营销的公司。这些公司拿到数据后进行剖析然后进行定向营销。

例如，NBA 球队菲尼克斯太阳队就是这项服务的客户之一。太阳队用它来找出观看比赛的人群住在哪里，以及了解观众赛后是否更有意愿光顾比赛的赞助商，从而加强其他地区的广告营销，如图 7-5 所示。

（2）精准营销（Precision Marketing）。提供广告投放支撑。

（3）移动商务（Mobile Commerce）。主要面向 Isis（Verizon、at&t 和 T-Mobile 发起的移动支付系统）。

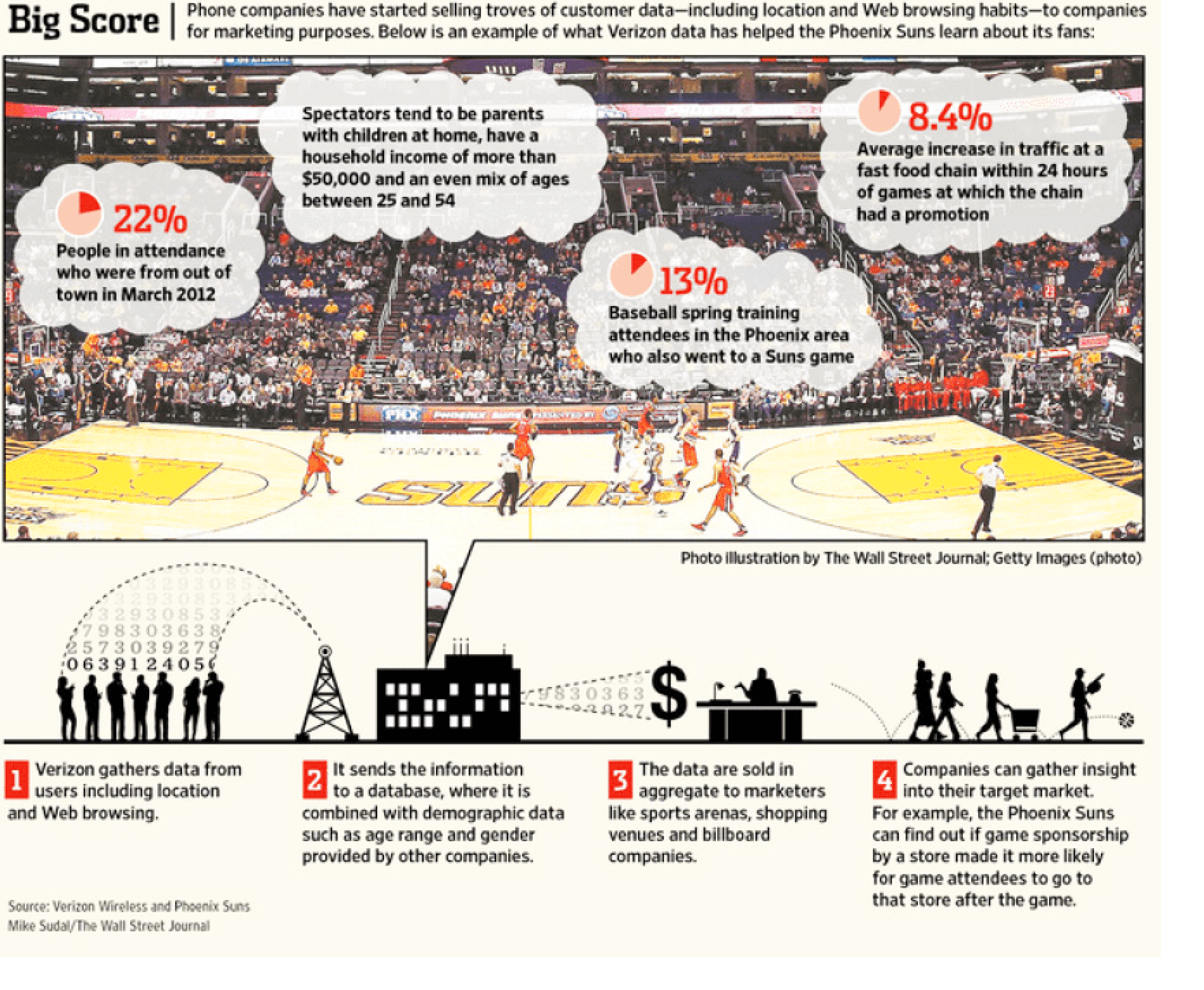


图 7-5 太阳队用 Precision Market Insights 分析商业数据

例如，美国的 Clear Channel Outdoor Holdings 是全球最大的广告牌公司之一，目前也在试用 Verizon 的 Precision Market Insights 服务。他们用这项服务来衡量开车经过广告牌的人看到广告后，有多少人会去商店购买广告产品。

【案例解析】：很长一段时间内，运营商在对外提供数据服务时，往往停留于提供原始数据层面，甚至违法违规事件屡有发生；而对于提供高附加值的数据分析服务，则是“雷声大，雨点小”，或者“说得漂亮，做的少”。

在本案例中，Verizon 成立了大数据部门，在运营商数据能力商业化方面迈出了可喜的一步。Verizon 通过更精准地掌握用户信息和用户行为，显然可以提高营销的定向性，如图 7-6 所示。在笔者看来，尽管运营商做的事情似乎跟水厂、电厂无异，但是其最大的不同正是在于管道里面的东西：数据流。跟管道流淌的水和电不同，运营商管道流淌的这种数据流绝对不是同质化的。通过对数据包的层层抽丝剥茧，是可以吸取出油来的。运营商只需对数据包进行深度分析，即可抓取 URL、关键字等信息。

专 家 提 醒

按照营销大师菲利普•科特勒的精准营销理论，“公司需要更精准、可衡量和高投资回报的营销沟通，需要更注重结果和行动的营销传播计划，还有越来越注重对直接销售沟通的投资。”

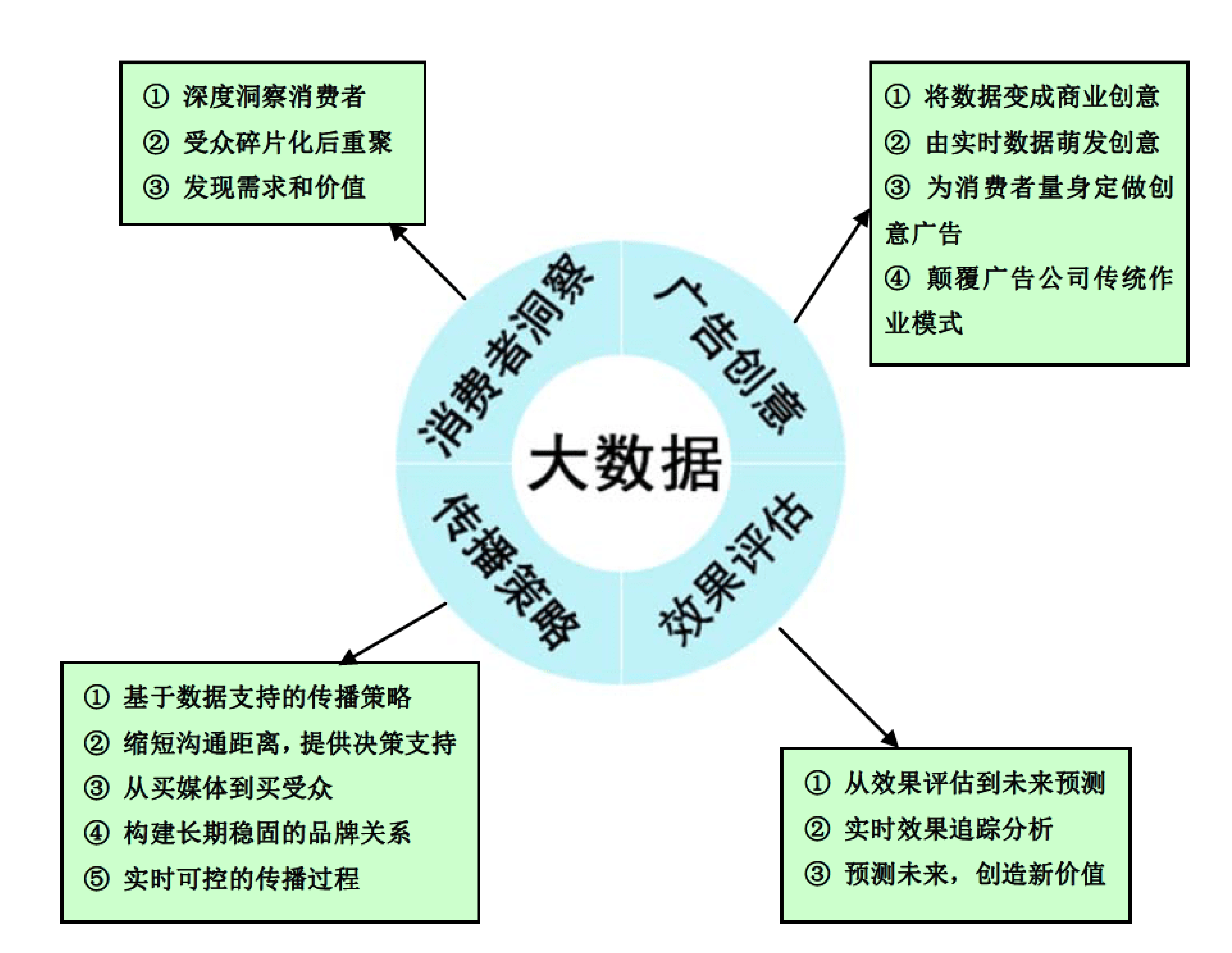


图 7-6　运营商在大数据时代的精准营销策略

7.2.4　【案例】中国联通开启大数据探索之路

据悉，中国联通在“移动通信用户上网记录集中查询与分析支撑系统”上引入了基于英特尔发行版 Hadoop 的大数据解决方案，并已经部署了 4.5PB 的存储空间，用于支撑全网数亿用户的查询工作。目前，该系统已经具备了每天处理 700 亿条上网记录的能力。

另外，中国联通目前正在着手对大数据业务进行研究，并已经成立了云数据运营中心，计划依靠该部门逐步尝试开展大数据业务的运营工作，并计划将该运营中心公司化，进行独立的运营，如图 7-7 所示。

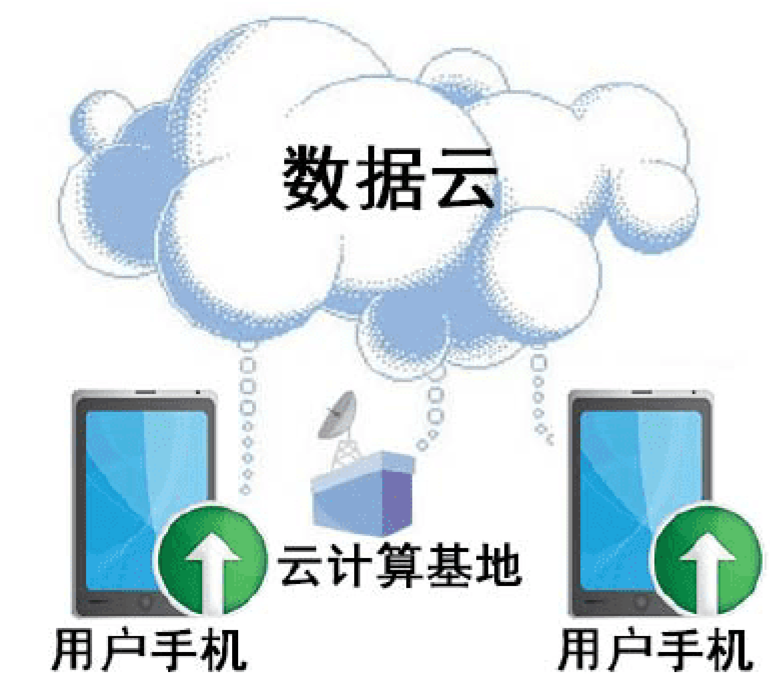


图 7-7　云数据运营

中国联通云计算基地选址在贵州省贵安新区电子信息产业园大数据核心区，计划投资约 50 亿元，主要建设基础构架、数据中心资源地、灾备系统、机房建设等设施。云计算基地项目建成后，将形成以云计算基地为基础、辐射周边的产业园区集群，带动战略性新兴产业全面、系统、有序发展，打造具备云计算基础的新兴产业聚集地。

中国联通研究院副院长黄文良表示，大数据业务开发的主要工作分为三步走：大数

据的采集、传输、集中存储；大数据的抽取、清洗、分类、挖掘、分析处理；基于大数据的业务和应用开发。

目前看来，大数据在中国联通的应用领域非常广泛，如应用于移动互联网、电子商务、BSS、市场营销、客户服务、网络建设维护等领域。大数据在中国联通的应用如表 7-2 所示。

表 7-2　大数据在中国联通的应用

主要应用	细节应用
上网记录查询	① 为移动用户的流量消费提供明明白白的清单 ② 为用户流量争议和投诉提供解决手段 ③ 提升公司服务水平，减少退费和赔付 ④ 为移动互联网用户上网行为、应用偏好分析提供基础信息
IP 地址溯源和上网日志留存	① 大数据系统可以满足国家对移动互联网不良信息的监管要求，为移动互联网的健康发展保驾护航 ② 对移动互联网的不良信息可实时监测和事后追溯
位置服务	① 基于服务用户手机的基站信息，可以获知用户的当前位置 ② 对用户的位置数据加以保存，可以实现用户全生命周期的轨迹服务。例如，在汶川地震和 2008 年雪灾发生时，可以找出当时全国联通有多少人漫游在发生地
短信漫游欢迎词	① 可全网统一漫游欢迎词的发送时间、发送内容、发送短信接入号、发送频次等标准 ② 漫游欢迎词可以作为公司自有业务宣传的窗口，也能开发出企业名片等后向收费业务，蕴含着巨大的经济效益
NET 取号及管道智能服务	① 智能手机上，用户都采用 NET 方式上网，用户手机号码的实时获取和传送，是实现移动互联网业务个性化服务的基础 ② 用户身份识别、终端、位置、承载网络等能力的开放，是智能管道的重要组成部分
内容计费服务	GGSN 设备（Gateway GPRS Support Node，网关 GPRS 支持节点）主要实现数据包在 WCDMA 移动网和外部数据网之间的路由和封装
3G 基站辅助规划和运行监测系统	① 利用周边已有基站的流量数据来预测新建 3G 基站的效益，可以提高新建基站的有效性 ② 可以实时掌握新建基站的竣工情况，并能对新建基站实现后评估 ③ 通过基站小区流量的异常变动数据，及时监测基站的运行情况
流量经营	① 利用大数据发现用户是 2G 终端还是 3G 终端 ② 利用大数据发现哪些 2G 基站下数据流量较高，分析其下用户是否是 3G 用户回落到 2G 基站下的？终端是否支持 3G 基站 ③ 通过大数据分析，得知用户偏好，把合适的应用推荐给合适的人，提高其数据使用量

续表

主 要 应 用	细 节 应 用
精准广告推送平台	① 搜索最能体现用户需求的内容，通过用户搜索内容可以了解用户的潜在需求 ② 对用户的搜索内容进行分析挖掘，可以实现针对用户需求的广告精准投放
终端管理服务	① 发现用户使用的终端类型，为用户应用针对性配置 ② 通过对用户使用终端历史的类型分析，可以发现用户的终端品牌偏好，实现新上市终端的定向推荐 ③ 通过实时分析新增终端的数据，可以实时了解终端的销售情况
客户互联网业务属性管理	发现和保存用户的互联网特征数据，这对新业务推荐和公司的流量提升等具有较高的价值

【案例解析】：Hadoop是个开源的系统，与一些商业系统比起来，成本是很低的；而且经过英特尔的“改良”和技术支持，使用者也能得到技术保障。在本案例中，作为电信运营商的主力之一，联通应该把握住这个环节，而现在主要的战略环节就是把握大数据的仓库。

笔者认为，联通作为电信运营商，没有必要跟其他的企业比拼，要做的事情是把大数据这座“金矿”管理好，并充分发挥其价值。同时，以“应用”为核心，通过对数据的深度挖掘、协同共享、应用整合，创新大数据产业发展模式。例如，运营商可以与互联网公司强强联合，构建先进的云平台，推出面向政府、行业、企业、公众的个性化应用产品，尝试合作运营增值服务，将云存储业务演变为“数据银行保险箱”业务，打造针对中小企业、行业用户的银行级数据存储平台。

7.2.5 【案例】法国电信大力发掘大数据价值

法国电信为了发掘大数据的价值，目前已在移动业务部门和公共服务领域进行了探索和尝试。

Orange Business Services是法国电信Orange的分部，同时也是法国最大的运营商，专门提供B2B（Business To Business，企业对企业之间的营销关系）服务，其拥有全球最大最畅通的语言和数据网络，覆盖220个国家及地区，其中166个设有当地支持，并提供云计算、企业移动性、M2M（Machine-to-Machine，即机器和机器的连接）、安全、统一通信、视频会议及宽带等综合通信服务。

Orange Business Services的策略是用云计算的方式为客户提供存储资源，使得企业客户能够以经济有效的方式妥善保存私有数据，并且充分发挥数据智能的作用。

在移动业务部门，Orange Business Services已在借助大数据改善服务水平，提升

用户体验。目前，法国电信开展了针对用户消费数据的分析评估，以帮助法国电信改善服务质量。

例如，当用户的通话突然中断时，Orange Business Services 会分析产生的原因并做出相应操作。除了技术故障外还有网络负荷过重，如果某段网络上的掉话率持续过高，则意味着该网络需要扩容。法国电信通过分析掉话率数据，找出了那些超负荷运转的网络，并及时进行了扩容，从而有效完善了网络布局，给用户提供了更好的服务体验，获得了更多的用户以及业务增长。

专 家 提 醒

Orange Business Services 虽然为客户提供数据存储系统，但是会严格遵守相关的隐私保护规定，不会去读取或者使用客户的这些数据。

另外，Orange Business Services 还承担了法国很多公共服务项目的 IT 系统建设，并在这些系统中开始尝试挖掘大数据的潜在价值。例如，Orange Business Services 承建了一个法国高速公路数据监测项目，每天都会产生 500 万条记录，对这些记录进行分析就能为行驶于高速公路上的车辆提供准确及时的信息，有效提高道路通畅率。

【案例解析】：在本案例中，Orange Business Services 目前已经能够提供涵盖 IaaS、WaaS（工作台站即服务）、SaaS 三个层面的“端到端”云计算解决方案。其中，大数据所需要的方案集中在 IaaS 层，Orange Business Services 在这一层面推出了以“灵活计算”命名的系列方案，突出使用灵活、计费灵活的特点，从而灵活满足用户对数据存储的　需求。

国外运营商已有一些突破性的应用案例，笔者觉得国内的运营商也应该紧抓这个机遇。对于运营商来说，大数据等于大价值。对于 IT 企业，大数据等于大机遇。通信行业需求从来都是 IT 技术发展的重要推动力，谁能得到通信行业客户的认可，必然会在大数据领域大有作为，进而成为大数据解决方案的领先者、领导者。

7.2.6 【案例】中国移动大数据全新战略定位

在 2012 年的移动互联网国际研讨会上，中国移动董事长奚国华提出了大数据时代全新的移动互联网战略，即构筑“智能管道”、搭建“开放平台”、打造“特色业务”与提供“友好界面”。这 16 字方针，体现了中国移动在移动互联时代全面开启之际的全新战略定位。

就中国移动的业务支撑能力而言，在业务量方面，用户总数超过 6 亿，全年受理营业 300 多亿次，支撑网连接了数十万台营业和客服终端，全年处理几万亿张计费话单，几千万张结算单，全网 OLTP 处理能力接近 40 亿 tpmc，存储的有效容量将近 20PB。这些数据都表明中国移动是一家名副其实的大数据的应用者。

针对企业客户的需求，中国移动以搭建平台、创新网络等方式，吸引更多产业链合作伙伴共同发展，打造现代信息服务产业链。大批设备制造商、系统集成商、内容服务提供商、营销代理商等集合到中国移动的支撑平台上，聚集起巨大的整合效应和能量，为企业客户提供基于大数据和移动互联网的信息化服务。

中国移动的经营分析体系所采取的是先构建数据仓库，再逐渐满足应用需求，即先把数据沉淀下来，再去考虑数据的使用问题。中国移动作为 IDC 业务新进入者，在竞争激烈的市场条件下，发挥出了其利用更先进的建造技术，进行更合理的布局规划的优势。

目前，中国移动已经在云平台上部署了分析型 PaaS 产品，利用 BC-Hadoop 构建大数据处理平台，并在英特尔“Xeon + Hadoop”平台上运行，同时建设了并行数据挖掘系统（BC-PDM&ETL）以及商务智能平台（BI-PAAS）等大数据应用平台，为将来进入大数据应用和服务市场做了充分准备。中国移动的大数据战略具体可以分为 3 步，如图 7-8 所示。

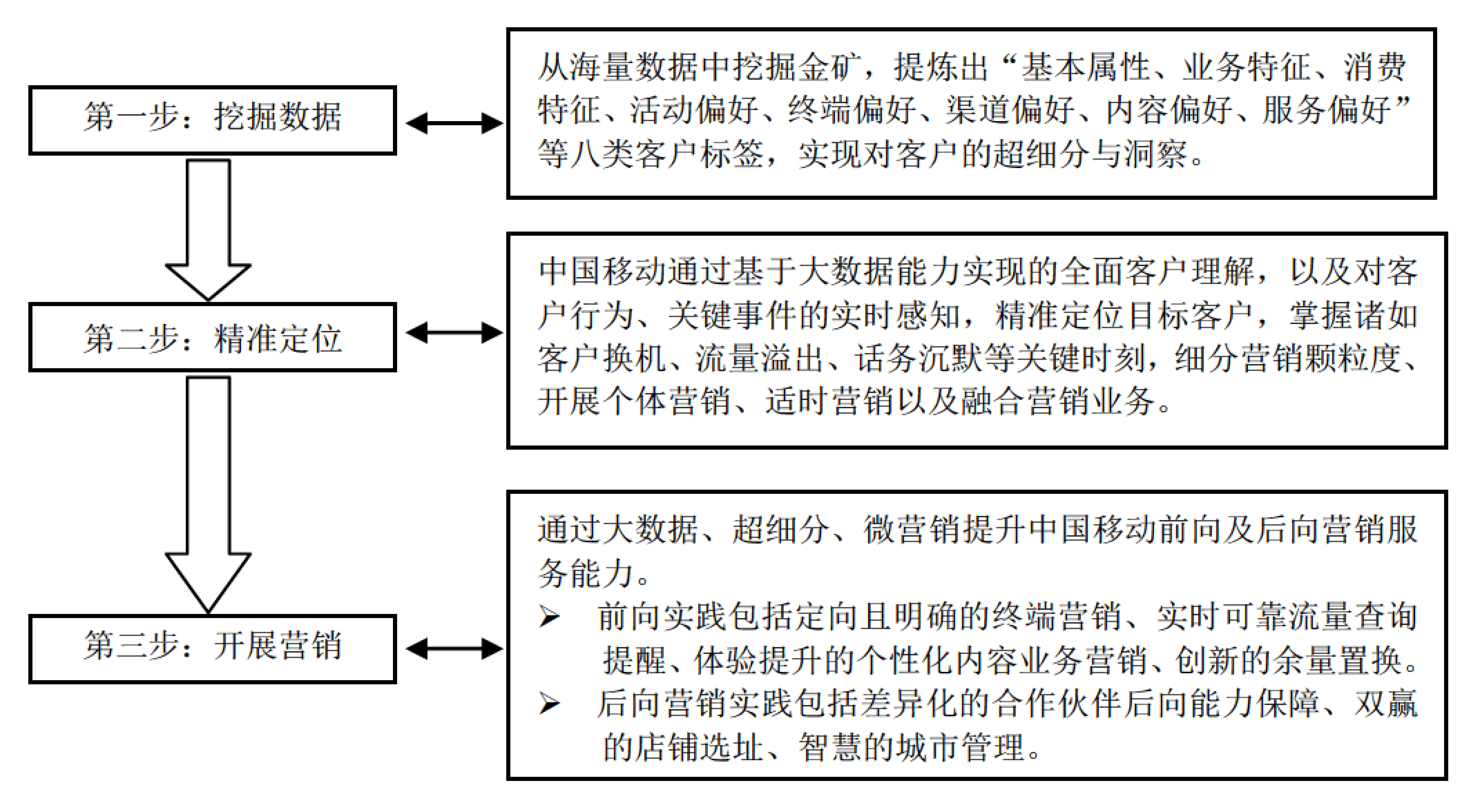


图 7-8　中国移动的大数据战略

目前，中国移动企业信息化系列产品已经得到 270 万家企业客户的认可，广泛应用于金融、交通、物流、IT、制造等领域，成为助力企业高效运作、引导大客户不断创新、推进中小企业快速成长、携手各方合作伙伴共赢的强大动力。据悉，中国移动在未来三年内还将投入超过 100 亿元资金，将 IDC 总面积扩容 6 倍，并引入全部主流的互联网服务商。

【案例解析】：在云计算、物联网等技术的带动下，中国移动互联网也已经步入“大数据”时代。如何利用这些“大数据”，从而开发出其中的价值，以及“大数据”将带来哪些领域的繁荣，成为运营商首要解决的问题。在本案例中可以看出，中国移动在满

足企业客户信息化需求上，正在逐渐形成一套成熟的信息服务产业链。

根据大数据数据量大、时效性要求高、数据种类及来源多样化等特征，运营商可以首先获取更多有用的大数据资源，例如，很多的网络运行信息，包含大量有价值的用户行为和位置信息，这样的信息可以加以利用。笔者认为，运营商有了资源就应该加以利用，避免大数据资源的浪费。

专家提醒

真正实现精准化营销和精细化运营的秘诀就在于如何利用好运营商手中的大数据。例如，移动掌握的海量话单、信令、互联网数据本身就是一笔宝贵的财富，利用好这些数据，充分、及时地对这些数据进行深度分析挖掘，不仅可以进一步提升服务质量、提高客户忠诚度、挖掘新商机、增加收入，还可以通过优化资源配置、减少浪费来提升运营效率，有效降低运营成本。

7.2.7 【案例】中国电信大数据聚焦商业模式

2009 年，中国电信启动了名为“天翼云”的云计算计划。

2010 年，中国电信便开始在全国 6 个省市部署了各种资源池（ResourcesPool），进行内部小规模的、商用实验和部署。

2011 年，中国电信也率先发布了专业化的运营思路，提出了专业化运营思路。

2012 年 2 月，中国电信首先成立了专业化的云计算公司。

截止到目前，中国电信对外提供了主机存储、CDN 等基础的云计算产品，并于 2013 年 6 月 1 日上线了云主机网上的实时销售，在政务监管、民生、医疗等领域提供了云产品的服务、云解决方案。

中国电信作为业内最大的数据中心服务提供商，目前在国内拥有近 300 个数据中心、5 个海外数据中心以及 4 个全国核心云数据中心。随着云产品的深入应用，中国电信试图探索大数据的商业模式。中国电信最有价值的大数据应用表现在 4 个方面，分别是语音数据分析、视频数据分析、网络流量分析、位置数据分析，如图 7-9 所示。

另外，中国电信提出了“智慧城市”发展战略，其中很重要的技术结合点就是物联网和大数据。在“流量经营”方面，中国电信从“话务经营”向“流量经营”转型。结合大数据技术，中国电信也将深入 IDC 服务以及智慧城市建设，并发掘移动互联与之结合的商机，重塑转型之路。

【案例解析】：总体来看，未来电信市场的一个重要方向是运营商将利用大数据来推动业务转型。这样电信业必将大部分的投资转向大数据应用市场。目前电信行业硬件增速较慢，但以云计算和大数据为代表的软件和服务已成为电信业 IT 投资的亮点，如图 7-10 所示。

语音数据

利用大数据处理平台分析呼叫中心的海量语音数据，建立呼叫中心测评体系和产品关联分析，可为保险公司等提供基于自动语音识别的大数据分析系统。

视频数据

基于智能图像分析能力的视频索引、搜索、摘要服务，从海量视频中挖掘有价值的视频信息，提供公用视频图像分析，中国电信全球眼智能系统在智慧城市、平安社区、交通监管等领域大规模地使用。

网络流量

通过分析互联网流量及协议信息，对一般性网络使用者的行为习惯分群组提供有针对性的网络便利性服务，例如精准广告。

位置数据

通过 LBS 系统平台，对移动通信使用者的位置和运动轨迹进行分析，实现热点地区的人群频率的概率性有效统计，例如根据景区人流进行基站优化。

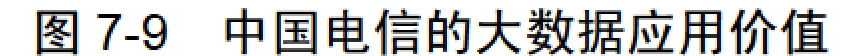


图 7-9　中国电信的大数据应用价值

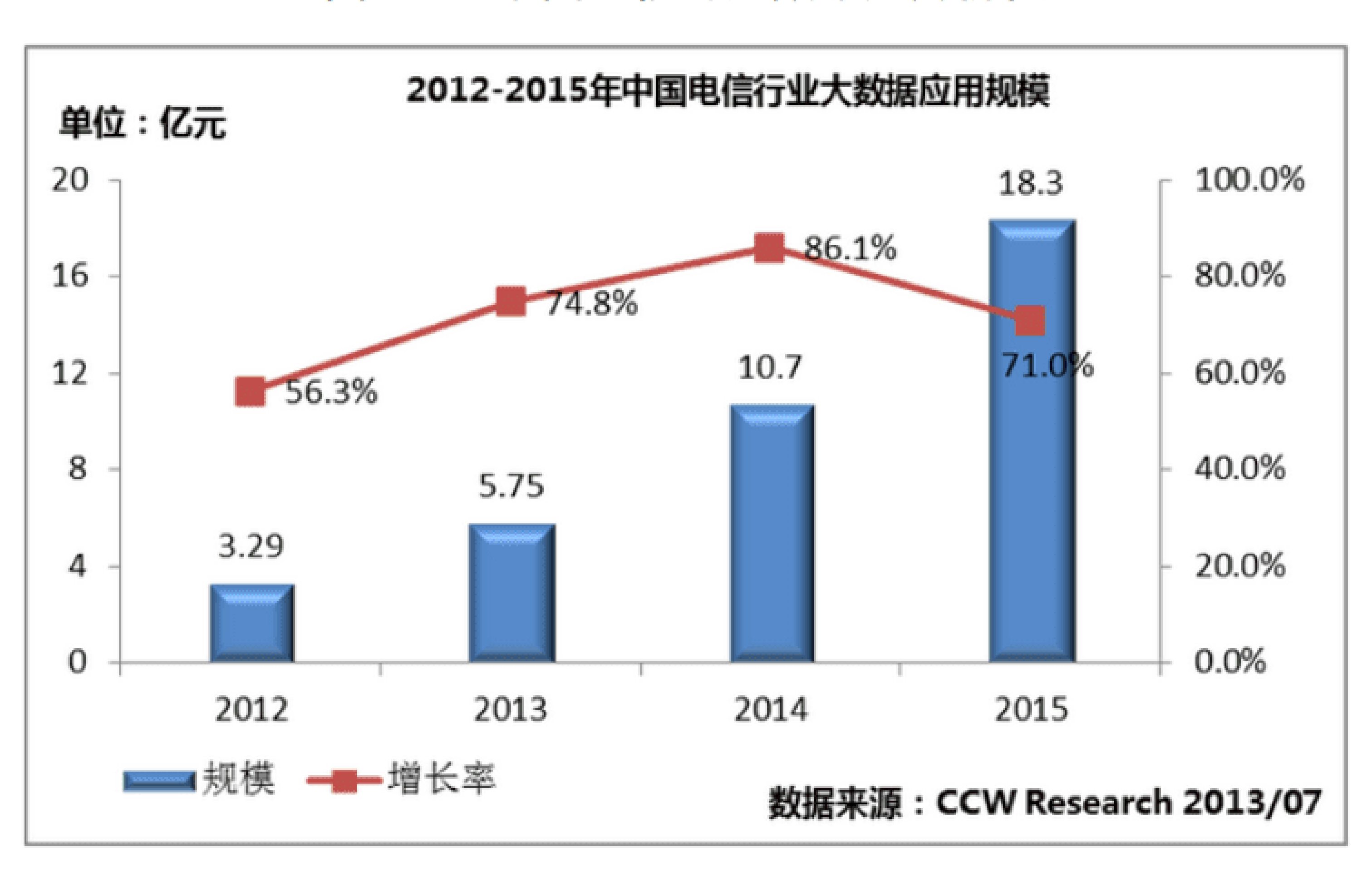


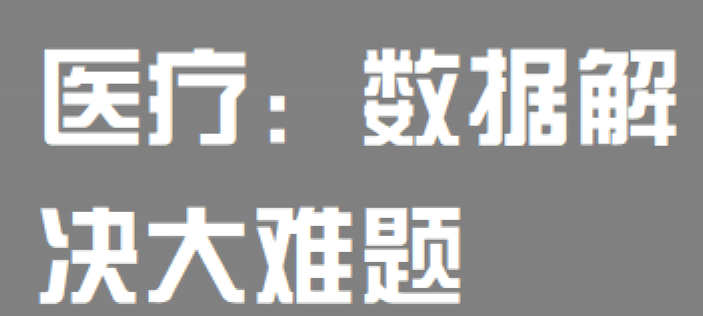
图 7-10　中国电信行业大数据应用规模分析

中国电信的数据只有通过长期的运营、使用和剖析后，才能够发挥出它的价值。在本案例中，笔者认为中国电信在做好数据挖掘和应用的基础上，将来还可以往前迈一步，帮助其他的中小企业，帮助需要这些服务的客户来提供一些数据的挖掘、平台和技术，这也许是电信运营商的机遇所在。

专 家 提 醒

云计算技术在数据中心领域是一个革命性的技术，对整个数据中心的发展有着重大影响。云计算模式可以动态扩展，并且可通过虚拟化资源、互联网方式来对外提供，政府和企业可以利用云计算的技术和资源来进行灵活、低成本、协同的 IT 应用部署。

8



医疗：数据解决大难题

学前提示

如何应对“大数据”，是摆在医院 IT 部门面前的一个“大考验”。如果处理不好，“大数据”就会成为“大包袱”、“大问题”；反之，如果应对得当，“大数据”则会为医院带来“大价值”。而这一切，都离不开科学地规划和部署存储架构。

要点展示

- 医疗行业大数据解决方案
- 医疗行业大数据应用案例

8.1 医疗行业大数据解决方案

随着大数据在医疗与生命科学研究过程中的广泛应用和不断扩展，其数量之大和种类之多令人难以置信。例如，一个 CT 图像含有大约 150MB 的数据，而一个基因组序列文件大小约为 750MB，一个标准的病理图则大得多，接近 5GB。如果将这些数据量乘以人口数量和平均寿命，仅一个社区医院或一个中等规模制药企业就可以生成和累积达数个 TB 甚至数个 PB 级的结构化和非结构化数据。

通过医疗大数据搜索病人信息，找寻疾病线索；通过移动 APP，市民与医生可以随时随地在线联系；通过物联网技术，病人个体化自我监测变成现实……近年来，信息技术在快速改变着传统医疗行业。大数据时代，以数据为内容的移动医疗会否颠覆传统医疗模式？它在医疗资源整合、医患关系改善方面又会有什么作为？

8.1.1 大数据在医疗行业的应用场景

医疗行业很早就遇到了海量数据和非结构化数据的挑战，而近年来很多国家都在积极推进医疗信息化发展，这使得很多医疗机构有资金来做大数据分析。因此，医疗行业将和银行、电信、保险等行业一起首先迈入大数据时代。麦肯锡在其报告中指出，排除体制障碍，大数据分析可以帮助美国的医疗服务业一年创造 3000 亿美元的附加价值。

专家提醒

医院和医疗行业面对的大数据主要有医学影像、视频（教学、监控）及文献等非结构化数据。由于这些数据增长很快且结构复杂，给数据管理和利用带来了较大的压力，存储与管理成本不断提高，数据利用困难且利用率低。

如表 8-1 所示，列出了医疗服务业 5 大领域（临床业务、付款/定价、研发、新的商业模式、公众健康）的 15 项应用，这些场景下，大数据的分析和应用都将发挥巨大的作用，从而提高医疗效率和医疗效果。

表 8-1　大数据在医疗行业的应用场景

5 大领域	应用场景	具体作用
临床操作	比较研究效果	通过全面分析病人特征数据和疗效数据，然后比较多种干预措施的有效性，可以找到针对特定病人的最佳治疗途径
	临床决策支持系统	临床决策支持系统可以提高工作效率和诊疗质量。目前的临床决策支持系统分析医生输入的条目，比较其与医学指引不同的地方，从而提醒医生防止潜在的错误，如药物不良反应
	医疗数据透明度	提高医疗过程数据的透明度，可以使医疗从业者、医疗机构的绩效更透明，从而间接促进医疗服务质量的提高

续表

5大领域	应用场景	具体作用
临床操作	远程病人监控	从对慢性病人的远程监控系统收集数据，并将分析结果反馈给监控设备（查看病人是否正在遵从医嘱），从而确定今后的用药和治疗方案
	对病人档案的先进分析	在病人档案方面应用高级分析可以确定哪些人是某类疾病的易感人群。例如，应用高级分析可以帮助识别哪些病人有患糖尿病的高风险，使他们尽早接受预防性保健方案
付款/定价	自动化系统	通过一个全面的一致的索赔数据库和相应的算法，可以检测索赔准确性，查出欺诈行为，避免重大的损失
	基于卫生经济学和疗效研究的定价计划	在药品定价方面，制药公司可以参与分担治疗风险，例如基于治疗效果制定定价策略。这对医疗支付方的好处显而易见，其有利于控制医疗保健成本支出
研发	预测建模	医药公司在新药物的研发阶段，可以通过数据建模和分析，确定最有效率的投入产出比，从而配备最佳资源组合。模型基于药物临床试验阶段之前的数据集及早期临床阶段的数据集，这样可尽可能及时地预测临床结果
	提高临床试验设计水平的统计工具和算法	使用统计工具和算法，可以提高临床试验设计水平，并在临床试验阶段更容易地招募到患者。通过挖掘病人数据，评估招募患者是否符合试验条件，从而加快临床试验进程，提出更有效的临床试验设计建议，并能找出最合适的临床试验基地
	临床试验数据的分析	分析临床试验数据和病人记录可以确定药品更多的适应症和发现副作用。在对临床试验数据和病人记录进行分析后，可以对药物进行重新定位，或者实现针对其他适应症的营销
	个性化治疗	通过对大型数据集（例如基因组数据）的分析发展个性化治疗，针对不同的患者采取不同的诊疗方案，或者根据患者的实际情况调整药物剂量，可以改善医疗保健效果，减少副作用
	疾病模式的分析	通过分析疾病的模式和趋势，可以帮助医疗产品企业制定战略性的研发投资决策，帮助其优化研发重点，优化配备资源
新的商业模式	汇总患者的临床记录和医疗保险数据集	汇总患者的临床记录和医疗保险数据集，并进行高级分析，可以提高医疗支付方、医疗服务提供方和医药企业的决策能力。例如，对医药企业来说，他们不仅可以生产出具有更佳疗效的药品，而且能保证药品适销对路
	网络平台和社区	网络平台和社区可以成为宝贵的数据来源，并产生大量有价值的数据。例如，Sermo.com 向医药公司收费，允许他们访问会员信息和网上互动信息
公众健康	大数据的使用可以改善公众健康监控。公共卫生部门可以通过覆盖全国的患者电子病历数据库，快速检测传染病，进行全面的疫情监测，并通过集成疾病监测和响应程序，快速进行响应	

8.1.2　如何从大数据中获取医疗价值

可以说，中国的医疗正在迈入“大数据”时代。医疗行业具有典型的“大数据”特征：一是数据量大；二是数据类型复杂。

因此，只有妥善处理好存储架构，“大数据”才能给医院带来“大价值”，才不会成为“大问题”。“大价值”的具体表现如图 8-1 所示。

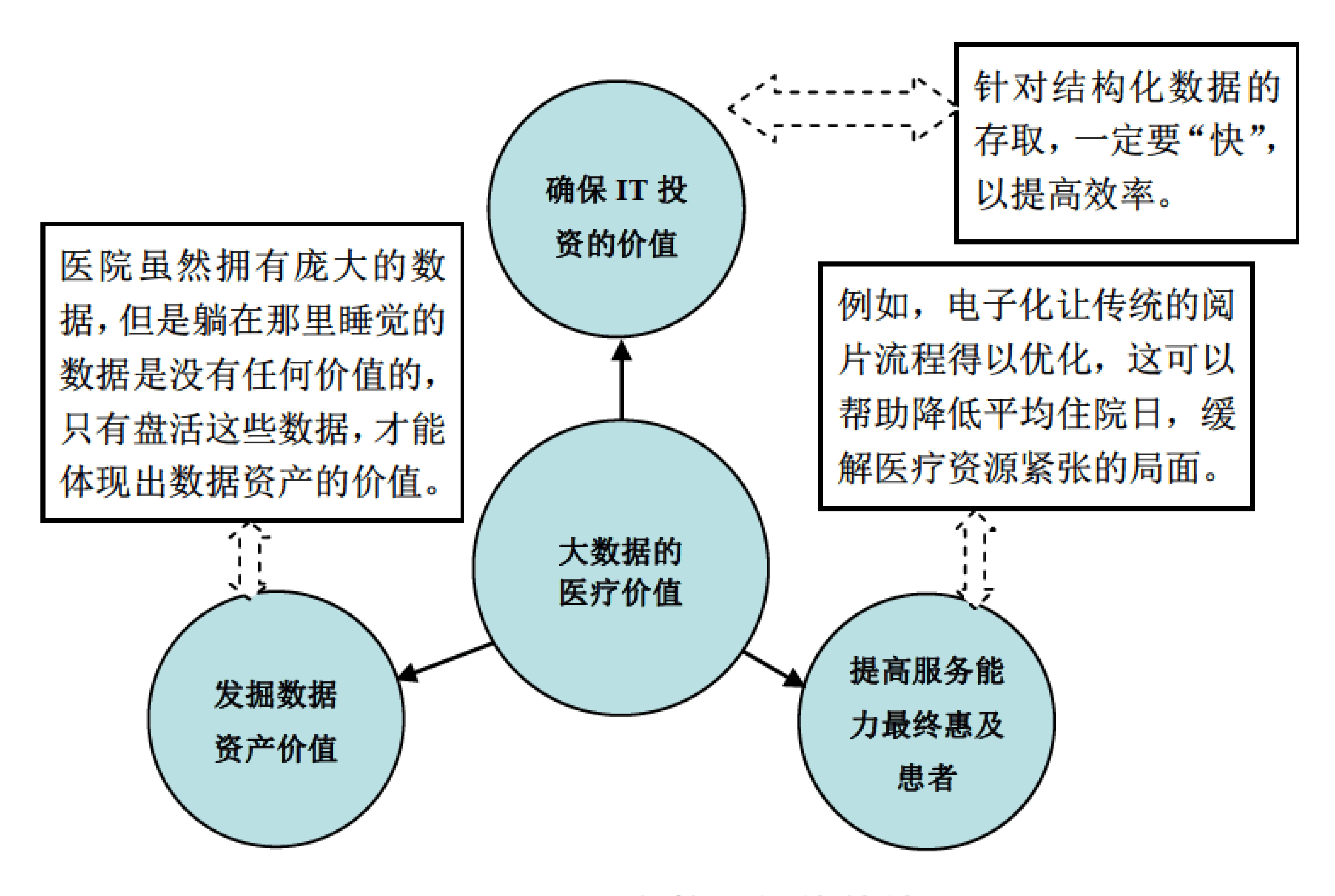


图 8-1　医疗大数据的价值体现

笔者相信终有一天，每个老百姓都可以随时管理、查询自己的健康医疗数据，不是在遥不可及的第三方，而是在他自己手里。而且这样的数据将不局限于体检结果、就诊记录，还可以延伸到你的基因数据，你的日常健康行为监测数据。你将从法律上拥有获得这些数据的权利！此时，我们可以真正地发挥医疗大数据的价值，人类对自身的认识也将上一个新的台阶。

8.1.3　医疗领域大数据的挑战和前景

大数据将成为行业和企业信息化建设的一道分水岭，擅用大数据，将会给信息化注入活力，并推动业务创新，最终帮助企业找到新的增长点；而错过大数据的发展机会，不但无法保证信息化建设的深入开展，也最终使企业丧失竞争优势。那么，在医疗领域，大数据又将面临哪些挑战？发展前景又会是怎样呢？

1. 大数据面临的挑战

面对“大数据”的挑战，医院必须考虑三个主要问题。

（1）数据存储是否安全可靠？因为系统一旦出现故障，首先考验的就是数据的存储、灾备和恢复能力。如果数据不能迅速恢复，而且恢复不到断点，则会对医院的业务、

患者满意度构成直接损害。

（2）如何提高医院运行和服务的效率？提高效率就是节省医生的时间，从而缓解医疗资源的紧张状况，这在一定程度上可帮助解决“看病难”问题。

（3）如何控制大数据的成本？存储架构是否合理，不仅影响到医院 IT 系统的成本，而且关乎医院的运营成本。医疗数据激增，造成医院普遍存在着较大的存储扩容压力。如今，医院的存储设备大多是来自不同厂商的完全异构的存储系统，这些不同的存储设备利用各自不同的软件工具来进行控制和管理，这样就增加了整个系统的复杂性，而且管理成本非常高。

专 家 提 醒

如何有效地将大数据存储成本降至最低，是企业和 IT 领导者，尤其是医疗大数据面临的根本性挑战。因为除了数据数量和形态的迅速增加，医疗数据还需要越来越长的保留期。患者的病历可能需要保存 70 或 80 年，甚至更长。许多情况下，病历还必须以原始格式永久保存，以满足法规的要求。

2. 大数据的发展前景

专家预测，至 2017 年，全球移动医疗市场价值将达 200 多亿美元，其中我国将占到三分之一。面对广阔的市场前景，怎样的移动医疗工具才会最终胜出？笔者认为，技术关键要链接医院、医生和病人，通过移动医疗让病人真正获益，医生收集数据后能有效改善医疗服务质量，只有做到这些，移动医疗才算两全其美。

2010 年，国家公布的“十二五”规划中指出要重点建设国家级、省级和地市级三级卫生信息平台，建设电子档案和电子病历两个基础数据库等诸项目标，也就是推进医疗信息化的“3521”工程，如图 8-2 所示。国家会逐渐加大对电子病历的投入，各级医院也将加大在数据中心、IT 外包等领域的投入。而随着医疗信息数据的几何倍数增长，医院信息存储将越来越受到重视，医疗信息中心的关注点也将由传统“计算”领域转移到“存储”领域上来。

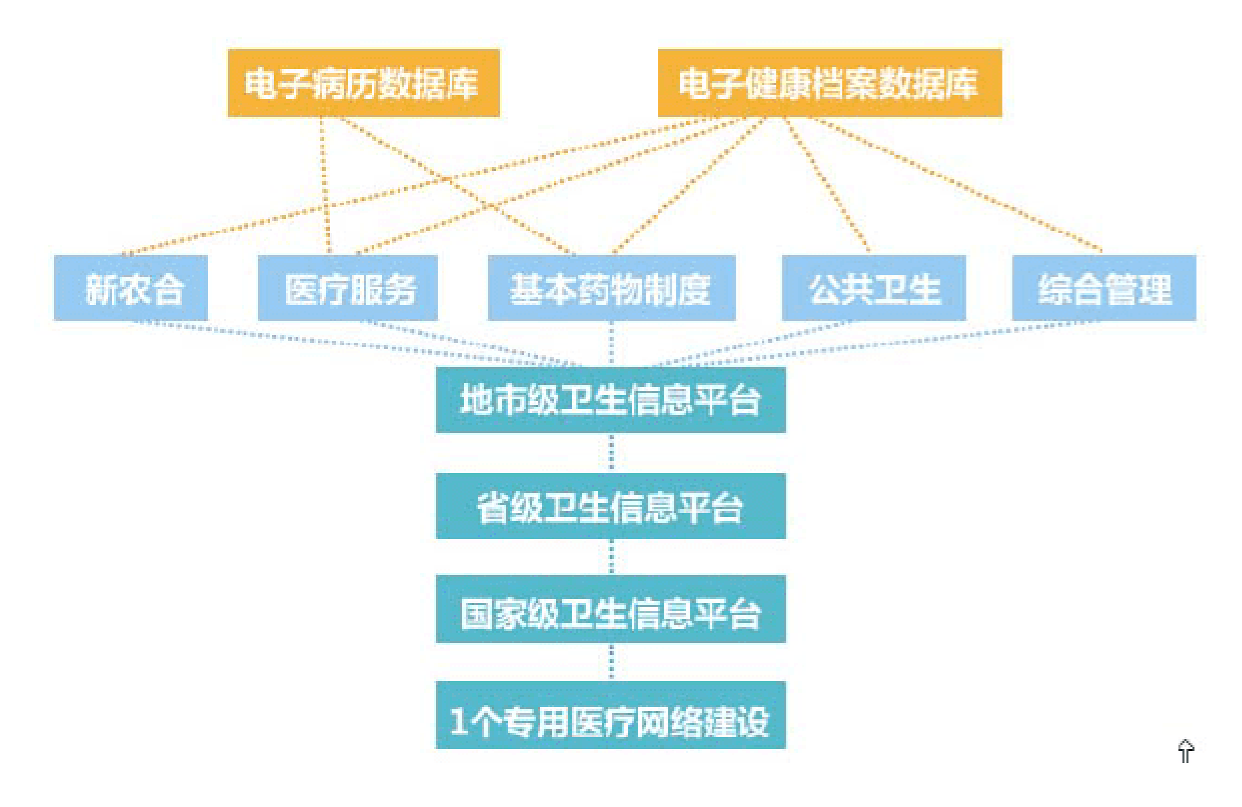


图 8-2 医疗信息化“3521”工程的基本构架

8.2 医疗行业大数据应用案例

如果说哪个行业从分析大量不同来源的数据中受益，那一定是医疗。在电子病历系统、图片系统、电子处方软件、医疗索赔、公共卫生报告、新兴的健康应用、移动医疗设备及医疗产业中，充满了等待被使用的数据。本节主要介绍信息医疗行业大数据的应用案例，希望对读者有一定的启发和学习价值。

8.2.1 【案例】利用大数据进行基因组测序

北卡罗莱纳大学（简称 UNC）在基因组测序技术上投入重资，以支持其医疗卫生系统更好地开展临床医护工作，同时推进基因组和生物基础研究。

该计划需要处理大量数据，要求管理和分析数百乃至数千人员的基因组，以满足临床医生和研究人员的不同需求。为了解决这种大数据难题，研究人员采用了三阶段流程，如表 8-2 所示。

表 8-2 基因组测序的主要工作流程

流程阶段	主要工作	细节说明
一阶段	在生物实验室中收集患者的组织	为每位患者生成数以亿计的短 DNA 序列，重新组合基因组并对重新组合进行质量控制，修正其间出现的错误
二阶段	检测个人的变异	使用大量的患者人群来帮助解决个人序列数据中的不确定之处
三阶段	向医生报告	收集了变异体之后，研究人员会在网站上将有关个人的信息提供给其医生

北卡罗莱纳大学的解决方案依赖于一个大型商业集群；该集群使用 50 个基于英特尔®处理器的刀片服务器，每周最多可处理 30 个基因组。目前，北卡罗莱纳大学在一个大型 EMC Isilon 数据系统上存储了大约 200～300TB 的基因组数据，如图 8-3 所示。利用 Hadoop 系统，研究人员可以进行极具针对性的分析，其很好地改进了 MapReduce 结构。



图 8-3 EMC Isilon 数据系统

专家提醒

刀片服务器是指在标准高度的机架式机箱内可插装多个卡式的服务器单元，是一种实现HAHD（High Availability High Density，高可用高密度）的低成本服务器平台，为特殊应用行业和高密度计算环境专门设计。刀片服务器就像"刀片"一样，每一块"刀片"实际上就是一块系统主板。

【案例解析】：在本案例中，基因组测序是一项新技术，各种事项都在迅速变化中。人们提出的问题也在迅速变化，因此信息解决方案也必须具有可调整性。

总体说来，大多数医疗机构的数据来自临床、财务、操作的应用程序。临床数据能提高医疗质量，使人口健康管理变得简单；财务数据帮助医院对盈亏底线做成本分析；而操作数据有助于设备管理和资源利用。把这些都综合在一起，就可以开始解决类似满足员工需求、提高工作效率和护理质量等大问题。

8.2.2 【案例】利用大数据来预防流感疫情

最近，美国波士顿和纽约宣布出现流感疫情。在波士顿市，目前已经呈报了 700 个案例，其中 18 人已经死亡。为了让疫情得到有效的控制，卫生官员以及应用开发人员向大数据寻求帮助。

虽然医生是控制疫情的"主战武器"，但是问题在于，目前并没有足够的疫苗可以普及所有的人群。此外，在研制流感疫苗之前，需要确认不同的流感病毒株，这样生产出来的疫苗才能真正防止流感的扩散。

因此，美国疾病预防控制中心（Centers for Disease Control，CDC）为了防止流感疫情的扩散，已逐步使用大量的数据来了解疫情。通常情况下，想要用流感疫苗阻止流感的蔓延，就需要精确地找到目前影响某个地区的流感菌株。CDC 通过对流感和肺炎死亡的跟踪，来了解流感疫情会不会造成死亡率上升。同时，CDC 也做了一些反病毒的耐药测试，用以确保流感疫苗可以缓解流感的影响。

与此同时，美国公共健康协会与斯科尔全球性威胁基金进行合作，推出了一款应用程序——FluNearYou，用于收集流感症状的发展信息。只要年满 13 岁周岁，都可以在网站上进行注册，该网站用以监测流感的蔓延程度，如图 8-4 所示。

专家提醒

FluNearYou 每周都会做一次调查报告，以帮助防灾组织、研究人员以及公共卫生官员为流感疫情的扩散做好准备。更重要的是，该数据共享应用程序对预测未来任何有可能的流感疫情爆发，都有极大的帮助。

作为全球最大的搜索引擎，每时每刻都有上百万用户在使用谷歌提供的搜索服务，其中搜索健康信息的人亦不在少数。这些用户行为提供了海量的有宝贵价值的分析数

据，当然对预防流感也是有重大意义的。

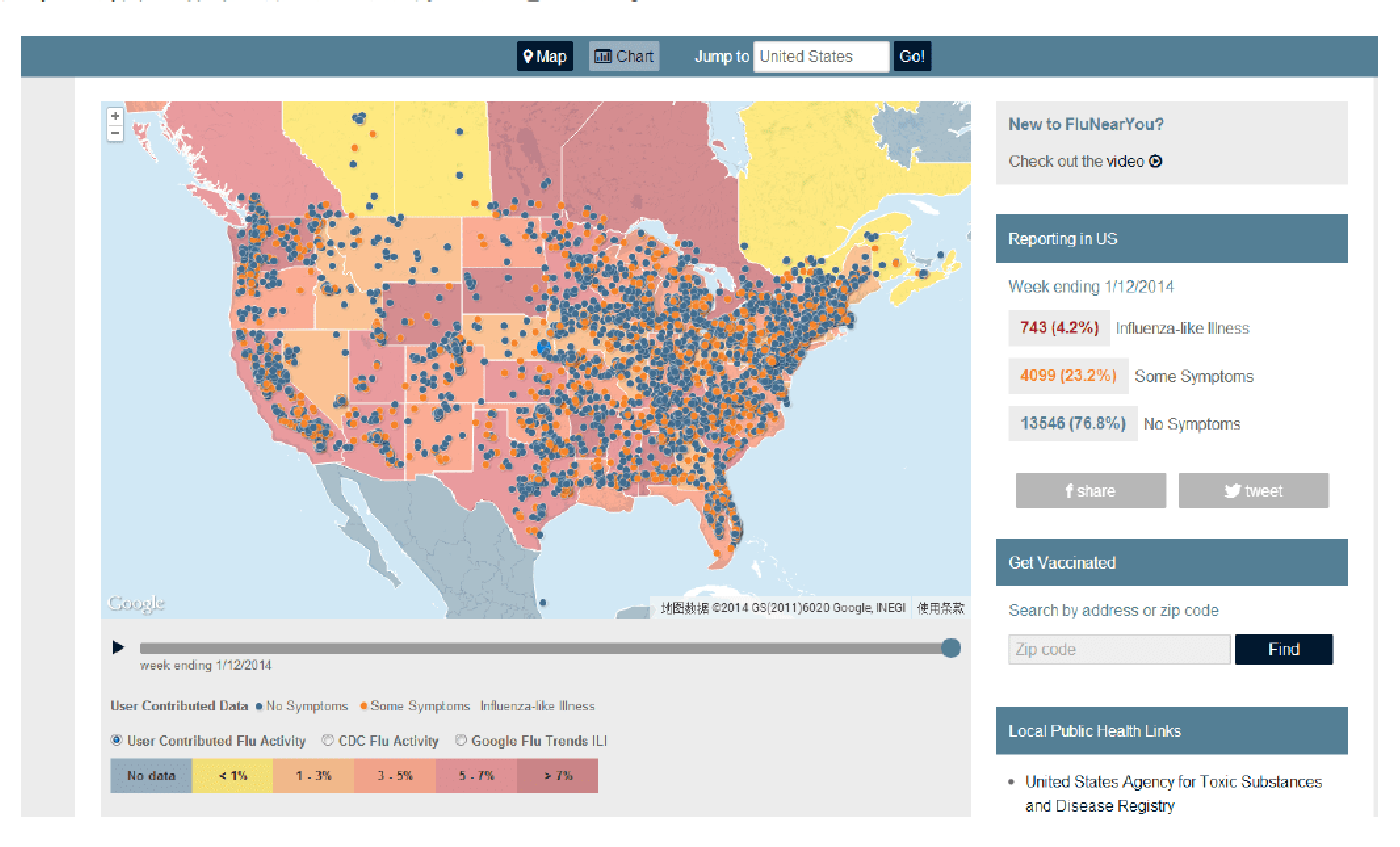


图 8-4　FluNearYou 主页上的流感地图数据

因此，谷歌开发了一款流感追踪器 Flu Trends，它可以监控相关的流感搜索字样，进而展示出在美国不同州的流感活动。美国疾病防止中心（CDC）是谷歌 Flu Trends 的研究合作伙伴。疾病预防控制中心的地图也能够显示流感疫情的扩散程度，如图 8-5 所示，这些数据将为人们提供流感早期警告。

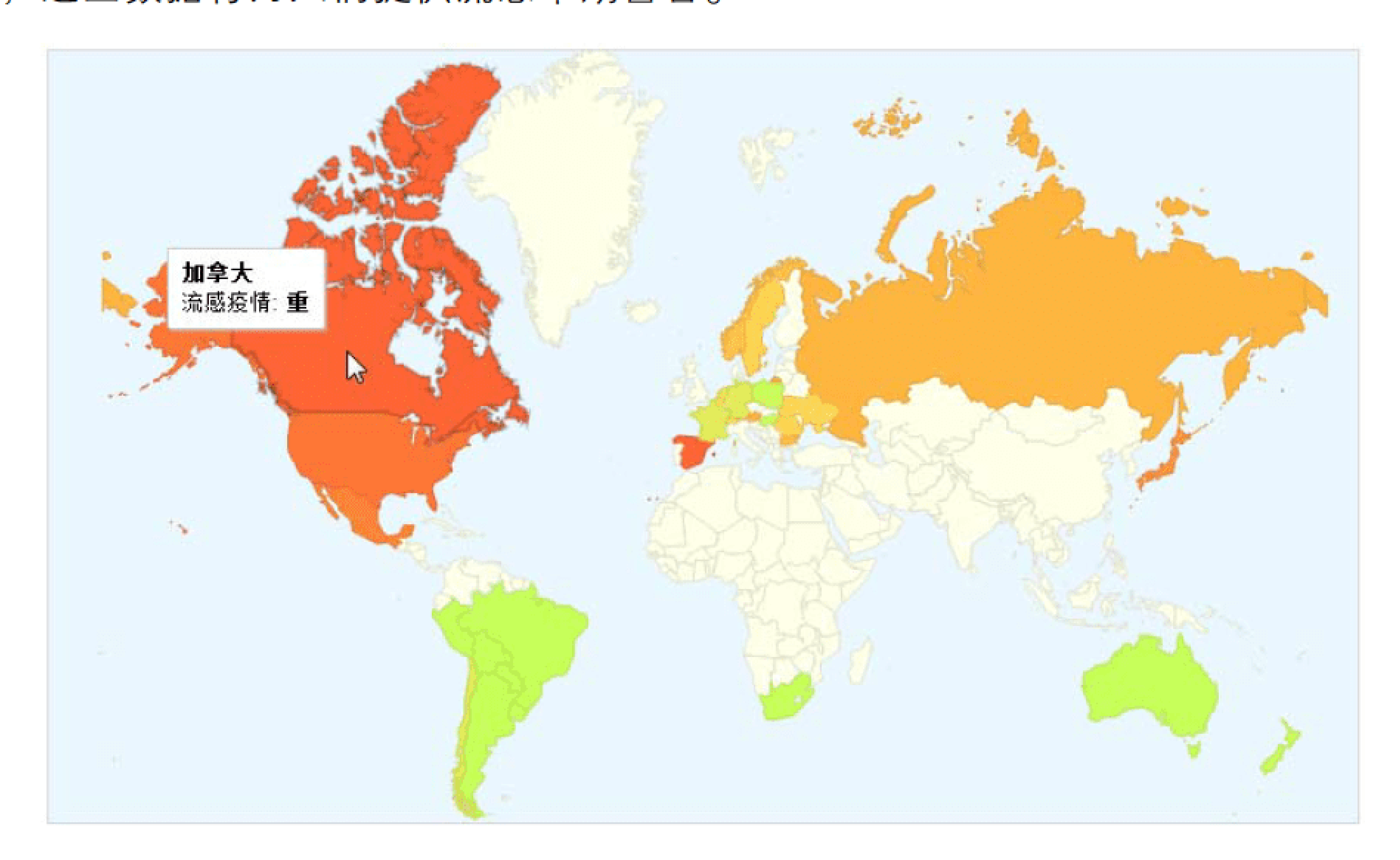


图 8-5　谷歌流感动态追踪地图

同时，谷歌还推出了 Flu View，也是一个跟踪工具，它接收并处理来自医生、医院以及 CDC 实验室的大量数据，为流感疫情的蔓延提供了一个清晰的图像，进而可以帮助医生能够有效地阻止流感疫情的蔓延。

目前，Google Flu Trends 已推广到全球 29 个国家，并由检测流感拓展到检测另一种感染性疾病登革热。在 Google Flu Trends 的启发之下，很多研究者试图利用其他渠道（例如社交网站）的数据来预测流感。

专家提醒

例如，纽约罗切斯特大学的一个数据挖掘团队就曾利用 Twitter 的数据进行了尝试。利用团队开发的文本分析工具，研究者在一个月内收集了 60 余万人的 440 万条 Twitter 信息，挖掘其中的身体状态信息。最终的分析结果表明，研究人员可以提前 8 天预报流感对个体的侵袭状况，而且准确率高达 90%。

【案例解析】：近些年，一些大规模的传播疾病一直没有间断，从非典到 H7N9，病毒性流感一波又一波袭扰人类，流感病毒不断变异并传播开来，令药物和疫苗要么准备不及，要么无法预防。但是如果能提早发现流感的发病趋势，不仅能为抗病毒药物的准备争取宝贵的时间，而且还有助于疫苗研发机构尽早采取措施。

可以想见，流感流行季，搜索流感症状的人会飙升，而在流感高发地带，这一比例会相应提高。这意味着流感相关关键词的搜索趋势与流感的流行趋势及严重程度存在某种程度的相关性。尽管并不是每个搜索这类关键词的人都有流感症状或患有流感，但把这些搜索结果汇总到一起时，或许可以从中建立起一个准确可靠的模型，实时监控时下的流感疫情，并对未来疫情状况进行估测。

本案例中的 FluNearYou 与 Google Flu Trends 都是采用这一大数据应用，来达到预测未来疫情状况的目的。其实针对美国在流感疫情防治领域所做的工作，中国疾病预防控制中心以及有关部门也可以学习，一个良好的疾病疫情监控信息系统，真的可以帮助控制疫情的蔓延，为我们的治疗防治工作赢得更多的时间。

不过，需要注意的是，即使在大数据的帮助下，医生永远也不可能完全地阻止流感的产生，医生能够做到最好的就是——控制流感疫情。

8.2.3 【案例】用大数据预测心脏病发作率

麻省理工学院、密歇根大学和一家妇女医院创建了一个计算机模型，可利用心脏病患者的心电图数据进行分析，预测在未来一年内患者心脏病发作的几率。

通常情况下，医生只会花 30 秒钟来观看用户的心电图数据（如图 8-6 所示），而且缺乏对之前数据的比较分析，这使得医生对 70%的心脏病患者再度发病缺乏预判，而现在通过机器学习和数据挖掘，该模型可以通过累积的数据进行分析，发现高风险指标。

【案例解析】：从本案例可以看到，将“大数据”运用到医学上不仅可以建立完善的医疗系统，更重要的是对于患者病情的预测以及控制会有巨大的作用。大数据一直在改变历史进程。而对于我们普通人而言，虽然对于大数据的概念云里雾里，但在生活中却每天都和它打交道。大数据也在不经意间改变着我们的小生活。



图 8-6　心电图数据

8.2.4　【案例】大数据 BI 促进医院智能化

近日，悉尼西区健康服务中心应用 BI 系统，使医院管理人员可以在几分钟甚至几十秒之内看到医院的各个环节的运行状况和管理状态，以及各个病人的状态如何、医疗服务如何等。悉尼西区健康服务中心所应用的 BI 系统具备三个特点，如图 8-7 所示。

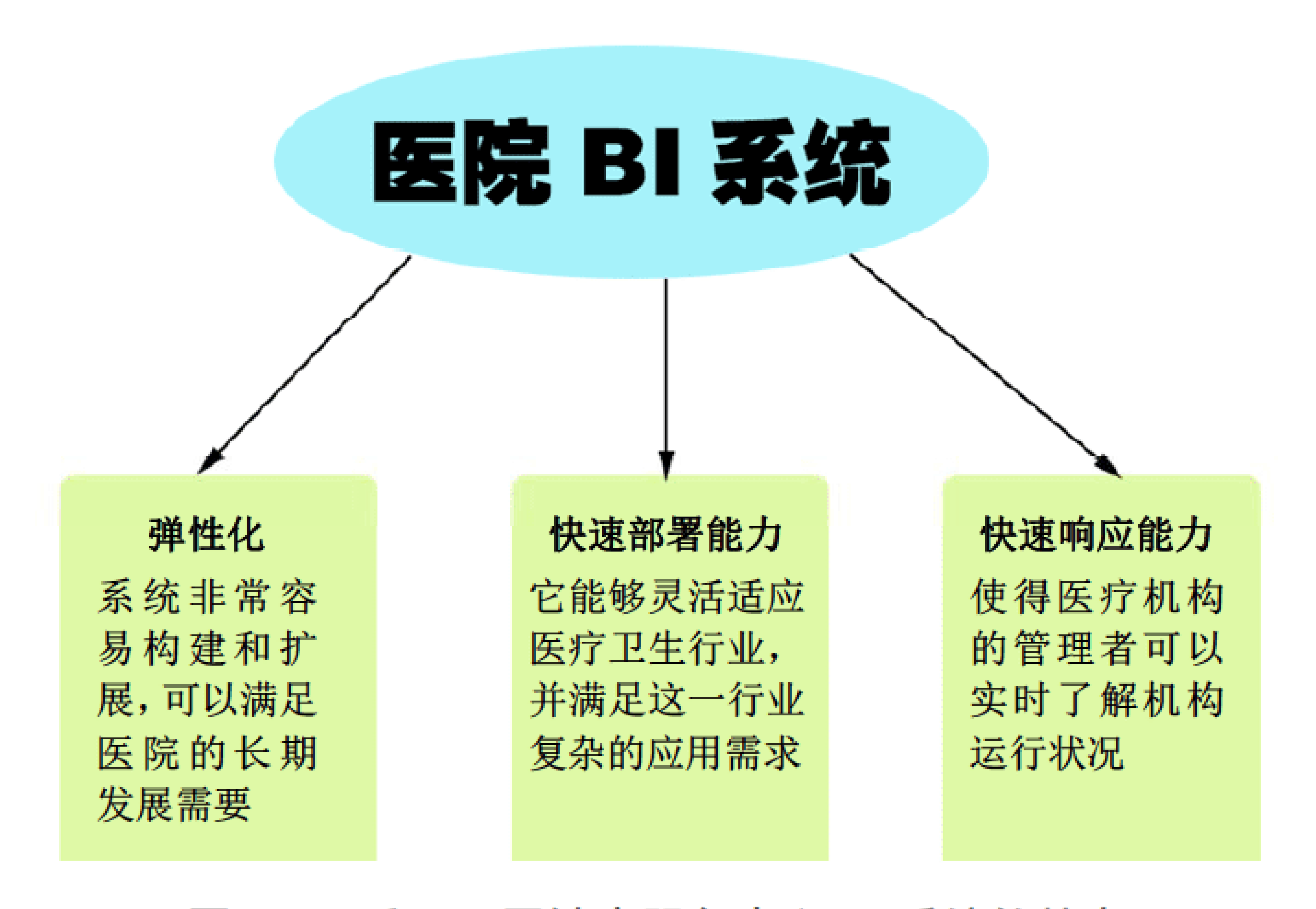


图 8-7　悉尼西区健康服务中心 BI 系统的特点

当然，并非所有的医疗机构应用了 BI 都能达到这样好的效果，一开始，悉尼西区健

康服务中心选用了 SAP 的 BI 产品，并在此技术上进行了二次开发，经过多年的发展，该中心终于使得 BI 切实融入到了整个 IT 架构中，并发挥出良好的作用。BI 的应用改变了传统的数据获取和分析方式，使得决策者可以通过快速准确的数据进行准确有效的决策。BI 不只是一种工具，它带来的是一种管理理念和手段的变革。

专 家 提 醒

SAP 是全球知名的企业管理和协同化商务解决方案供应商，其致力于为企业实现卓越运营提供领先的企业应用云计算、商务分析、移动商务、内存计算等解决方案。SAP 大数据解决方案主要集中在数据库及数据仓库层面和企业信息管理层面。其中，数据仓库及数据仓库解决方案主要由实时数据平台 HANA、分析型数据库 SAP Sybase IQ 和交易型数据库 Syabse ASE 来处理，企业信息管理主要由 SAP Information Steward、SAP NetWeave、企业内容管理（ECM）来处理。

【案例解析】：医疗行业是世界上最复杂的行业之一，因为在医疗机构中，它所服务的对象是各种不同类型的人，这里不仅包括提供服务的医生、护士，还包括不同类型的患者，再加上医院的基础设施、各种医疗器械等都需要管理，这些都给医疗行业的运作带来了很大的复杂性。

大数据 BI 系统正是以上这些问题的最好解决方式。大数据 BI 是能够处理和分析大数据的 BI 软件，区别于传统 BI 软件，大数据 BI 可以完成对 TB 级别数据的实时分析。

例如，国内很多医疗机构非常热衷于采购医疗器械，如 CT、核子共振等高级设备，应用这些设备确实能够提升医院的服务能力，医院也能借此获取更多的收益。但是，如果这些设备中所产生的数据无法快速传达到医生那里，供他做出参考和判断，势必会大大降低设备的效率，设备本身的价值会被浪费掉。目前，大部分医疗器械都是数字化产品，它们的应用都需要与之相配套的 IT 系统作为支撑，以便让其产生的数据能够快速传递出去，才能真正发挥其作用。

笔者认为，在 BI 系统的应用上，医院应该以现有的成熟的 BI 产品为基础，进行一些自己的开发，并将 BI 系统与其他医疗信息化系统整合起来，这样才能发挥其作用。此外，对于那些专业的医疗信息化系统，医院没有必要自己开发，只需要选用成熟的产品，并在异构的系统上进行二次开发，将其集成在一起即可。

专 家 提 醒

如果说 IT 系统已经成为医院的“血液系统”和“循环系统”，那么，大数据 BI 已经成为医院的“神经系统”。

8.2.5 【案例】用大数据“魔毯”改善健康

不久前，英特尔（Intel）、通用电气（GE）联合宣布，两家公司已经达成最终合作

协议，共同出资成立一家新的医疗保健公司，关注远程医疗和独立生活。

医疗创新公司主要业务是开发和推广能够增强家庭和社区健康、独居生活的产品、服务和技术，并重点关注三大领域：慢性病治疗、独立生活、辅助技术。

医疗创新公司成立不久后便推出了两款针对家庭医疗的产品：

- Health Guide。Health Guide 适用于慢性病人，可以监控各种人体机能，提取吃药时间、血压、体重等数据并发给相关的医疗机构；它还支持病人和医生进行电话和视频会议，从而提升病人的生活质量，让病人不必总是亲自到医院看医生。Health Guide 产品如图 8-8 所示。
- Reader。它是一种便携式设备，可自动将印刷文本转换成数字文本并朗读出来，帮助盲人和有阅读障碍的人进行阅读，如图 8-9 所示。




图 8-8 Health Guide



图 8-9 Reader

目前，医疗创新公司正在研究一种“魔毯”，这块地毯配备传感器和加速器，可以安装在老年人家中。传感器可以感应那些缺乏人照料的老人下床和行走的速度和压力，一旦这些数据发生异常则对老人的亲人发送一个警报。

【案例解析】：当今一系列重大社会问题，包括人口老龄化、高昂的医疗成本、为数众多的慢性疾病患者等，需要新的护理服务模式来解决。笔者认为，我们必须跳出“去医院和诊所看病”这种旧模式，转变为以家庭和社区为基础的护理模式，从而将预防、早期诊断、医疗保健行为改变和社会支持结合起来。

在本案例中，虽然内置传感器装置对大多数人来讲依然昂贵，但由于这些将自身数据量化的小工具越来越受到欢迎，用户可以清楚地了解和改变自身的行为，从而改善健康状况。

8.2.6 【案例】用大数据分析找出治疗方案

代谢综合征（Metabolic Syndrome，MS）是多种代谢成分异常聚集的病理状态，

是一组复杂的代谢紊乱症候群，是导致糖尿病（DM）、心脑血管疾病（CVD）的危险因素，其集簇发生可能与胰岛素抵抗（IR）有关，目前已成为心内科和糖尿病医师共同关注的热点，国内外至今对它的认识争议颇多。

美国安泰保险为了帮助改善代谢综合征患者的预测，从一千名患者中选择 102 个完成试验。在一个独立的工作实验室内，通过患者的一系列代谢综合征的检测试验结果，在连续三年内，扫描 600000 个化验结果和处理 18 万个索赔事件。

安泰保险通过大数据分析，将最后的结果组成一个高度个性化的治疗方案，以评估患者的危险因素和重点治疗方案。

【案例解析】：大多数疾病可以通过药物来达到治疗效果，但如何让医生和病人能够专注参加一两个可以真正改善病人健康状况的干预项目却极具挑战。在本案例中，安泰保险正尝试通过大数据达到此目的。笔者也认为，让保险公司在先进的分析上花钱，比起让医疗机构来投资简单得多。

8.2.7 【案例】手表成为大数据的有力武器

据美国心脏学会说，每 4 个美国人中就有一人患高血压，这些人中还有三分之一的人根本未意识到。虽然每个医生都会对患者量血压，但是没有几个人会 24 小时监测病人血压。

近日，新加坡研究人员发明了一种名为 BPro 的黑色塑料血压监控手表，只要戴在患者的手腕上，就会 24 小时密切监控血压，如图 8-10 所示。BPro 内部有一个传感器，通过计算手腕上动脉跳动的次数，再转换成血压读数。BPro 除可显示波浪形曲线，表明心脏跳动频率和力度外，还可显示血压方面任何令人担忧的趋势。



图 8-10　BPro 血压计

人们在医院测量血压时，紧张的心情可能导致血压异常。此外，人体血压随时在发生变化，即使单独一次测量能够得出准确结果，也难以反映心血管系统运作状况的全貌。与需要暂时阻断动脉血流然后放气来测量血压的传统血压计不同，BPro 血压计通过监测脉搏波沿手部动脉的传播速度来计算血压，它还比一般的便携血压计轻便得多，可以像手表一样随身佩戴。

研究人员不仅用 BPro 治疗那些血压非常高的人，也正把目光瞄准那些没有任何症状的人。让病人戴上这种血压监控手表，不仅可能降低心脏病和中风发病率，还可收集大量数据。通过持续测量血压状况，BPro 使医生能详细了解佩戴者的血压变动，及时发现异常状况，最终将有可能利用这些数据来预测心脏病发病时间。

【案例解析】：从本案例可以看到，大数据的挑战不仅来自数据量的增长，还需要新技术的支持。因此，信息化如果和健康整合就会关系到每一个人的生活、健康，我们可以去展望，数据是“新的石油”，我们怎么找到这个能源和挖掘它，这是非常值得研究的。

专家提醒

笔者认为，大数据趋势下的大服务时代，用户与厂商都需要拥有主动意识，以最大化挖掘数据价值为目标，不能坐等应用需求。

8.2.8 【案例】中南大学启动临床大数据系统

2014 年 1 月 14 日，中南大学宣布该校“湘雅临床大数据系统建设项目”正式启动，首批共 101 个项目入选，覆盖 40 余个临床学科。据悉，开展大数据在临床医学领域大范围、系统性的探索和应用，这在国内高校中尚属首次。

中南大学所属的湘雅医院、湘雅第二附属医院、湘雅第三附属医院、湘雅口腔医院和湘雅医学院肿瘤医院每年门诊人次过千万，每年住院人次超过 35 万，手术人次每年至少是 20 万人次以上，医疗体量极为庞大，可产生海量的多媒体临床数据。如将其运用于临床科研和转化医学研究，进而带动基础医学发展，将有助于产生更多有价值的成果。

中南大学将分 5 年连续投入 1 亿元人民币，资助该校所有临床专科建设其大数据系统，并为每位受助医生配备一名软件专业研究生，协助开展数据采集，以建立起从病人踏入医院门槛开始的一整套网络化电子病历系统。

临床大数据系统的数据采集主要包括以下两方面的内容：

（1）基本信息。包括患者年龄、民族、职业、工作等基本信息；婚姻、月经及生育情况；家庭健康及疾病情况；生活、卫生情况；不良嗜好等信息。

（2）病历信息。包括患者的主要症状、体征及疾病发生时间等信息；疾病发生、发展及变化过程和诊疗信息；患者既往的健康及疾病信息；疾病诊断、个性化内科治疗和手术治疗等；治疗效果和药物反应等情况；疾病发展或痊愈情况；患者随访跟踪情况；患者生物标本储存及相关信息等。

同时，笔者还在现场看到，医生只需在手机上登录采集系统，点击体温、脉搏、血压等按钮，便可指挥与一位被测者相连的采集设备开始工作，并在手机上实时读到动态测量值。采集完成后，检测数据通过 WiFi 网络被发送至后台的大数据中心储存。如有其他人需调阅该数据，只需从另一台手机登入推送系统，便可收到大数据中心实时发送过来的完整记录。

未来，中南大学还将通过临床医学与信息技术的深度融合，深度挖掘和分析大数据，

将建立国际先进和国内领先的医疗相关数据运营模式，促进智慧医疗、个体化医疗、医院精细化管理、临床科研、转化医学和基础医学的发展，项目可以为卫生行政部门提供决策依据。

【案例解析】: 在本案例中，临床大数据系统的建立对诊疗模式变革意义重大。在血管外科手术日益精细和复杂并趋向个性化的今天，临床大数据系统不仅有助于医生提高诊疗和科研水平，对病人了解自身详细病史也极为有利，其必将对病因诊断、用药、手术、预后等产生积极而深刻的影响。

国民健康和医药卫生事业的发展是构建和谐社会的重要因素。国际上，有些大学和科研机构已经开始针对某个或少数的疾病进行有关临床大数据的研究，如美国匹兹堡医学中心设立了乳腺癌临床大数据库。但迄今为止，国内还没有开展大数据在临床医学领域大范围、系统性的探索和应用，希望此次中南大学可以带来一个好的开端。


读书笔记

9

网络：抓牢数据发源地

学前提示

巧妇难为无米之炊，大数据的关键在于谁先拥有数据。互联网提供了数据来源，数据分析能够针对每一位用户的信息做精准匹配。面对互联网的海量信息，数据的作用将远远超出以往。可以说，互联网推动了大数据由后台走向前台。

要点展示

- 互联网大数据解决方案
- 互联网大数据应用案例

9.1 互联网大数据解决方案

网络社交过程中，每天都会产生大量的数据，但是它们并不像我们想象中的那样是冷冰冰的、枯燥的数据，而是更加活生生的、有趣的数据。这些数据不同于以往单纯的数字，它们声色结合、图文并茂。

全球畅销书《社会消费网络营销》作者拉里·韦伯指出："所谓大数据，包括企业信息化的用户交易、社会化媒体中用户的行为、关系以及无线互联网中的地理位置数据。"大数据捕捉到了社交网络中"人"的踪迹，而智能广告则是利用数据追踪、研究、理解"人"，从而选择"对的人"与"对的时机"。

9.1.1 传统互联网大数据解决方案

互联网（Internetwork，Internet），始于 1969 年的美国，又称因特网，是全球性的网络，是一种公用信息的载体，是大众传媒的一种。互联网具有快捷性、普及性，是现今最流行、最受欢迎的传媒之一。互联网这种大众传媒技术，比以往的任何一种通信媒体都要快。互联网行业是"人"的网络消费，市场大是行业发展最重要的因素，腾讯等一批内地互联网企业的发展都受惠于此。

1. 传统互联网的盈利模式

目前，传统的行业门户网站的盈利模式主要由 4 大基点作为支撑，分别是广告盈利、会员盈利、活动盈利以及商务盈利，如表 9-1 所示。此外，笔者还注意到不少门户网站，由于不满足于原有既定的盈利模式，正在努力谋求新的利润基点，其中电子商务盈利是很重要的一个组成部分。

表 9-1 传统互联网企业的盈利模式

盈利模式	主要特点	面临问题
广告盈利	凭借广告谋求门户网站盈利，几乎是所有门户网站盈利模式的首要选择	依靠广告产生大规模的网站盈利，难度是很大的，只有极少数业内特别出色的门户网站可以做到
会员盈利	通过吸纳会员，收取会员费，从而使得网站产生利润，是目前已经被证明的比较切实可行的途径，如栖息谷、世纪佳缘、嫁我网等	需要网站本身在业内有一定的影响力，与网站广告如出一辙
活动盈利	通过策划活动扩张网站的影响力与知名度，同时谋求更强的盈利点，是所有门户网站运营的必由之路	需要线上与线下的双方互动，规模和成本难以控制
商务盈利	将门户网站与电子商务进行有机结合，是目前整个行业的新动向	数据量较大，难以管理

2. 传统互联网如何利用大数据

虽然大数据目前在国内还处于初级阶段，但其商业价值已经显现出来。手中握有数据的公司站在“金矿”上，基于数据交易即可产生很好的效益；基于数据挖掘会有很多商业模式诞生，例如帮企业做内部数据挖掘，或侧重优化，帮企业更精准地找到用户，降低营销成本，提高企业销售率，增加利润等。

那么，传统互联网企业该如何利用手中的“金矿”呢？笔者认为可以从网络广告、数据挖掘、数据分析以及实施决策 4 个方面入手，如图 9-1 所示。



图 9-1　传统互联网企业掘金大数据的方法

大数据将成为互联网时代的“发动机”，互联网不再只是媒体，更是用户不断转化的平台，而数据在营销全程中扮演的角色也必然要由参考工具转向驱动发动机。数据驱动的精准营销引擎，将颠覆传统的营销决策模式及营销执行过程，给网络营销行业乃至互联网及传统行业带来革命性的冲击。

专 家 提 醒

以阿里巴巴为例，2013 年阿里巴巴“双十一”的交易额达到 350 亿美元，超过内地日均零售总额一半。如此大的数据量和集中化数据处理，背后需要的是强有力的网络支撑平台。阿里巴巴搭建的先进可靠的数据中心，为“双十一”突增的数据量，提供了可靠的基础设施保障。不仅是阿里巴巴，京东商城为了更好地应对互联网化，提升竞争力，提出了“技术驱动”的口号，其技术的核心和内涵就是云计算和大数据，以利用大数据和云计算驱动京东在自营 B2C、开放业务和金融业务的发展。

9.1.2 移动互联网大数据解决方案

移动互联网，就是将移动通信和互联网二者结合起来，成为一体。移动通信和互联网成为当今世界发展最快、市场潜力最大、前景最诱人的两大业务，它们的增长速度都是任何预测家未曾预料到的，所以可以预见移动互联网将会创造经济神话。

如今随着智能手机时代的来临，移动互联网行业也在迅速发展。最近有消息表示，2013 年全球的移动互联网用户达 24 亿。2006—2015 年中国移动互联网市场规模如图 9-2 所示。另外，我国移动互联网用户还在不断地向传统互联网和手机用户渗透，如图 9-3 所示。

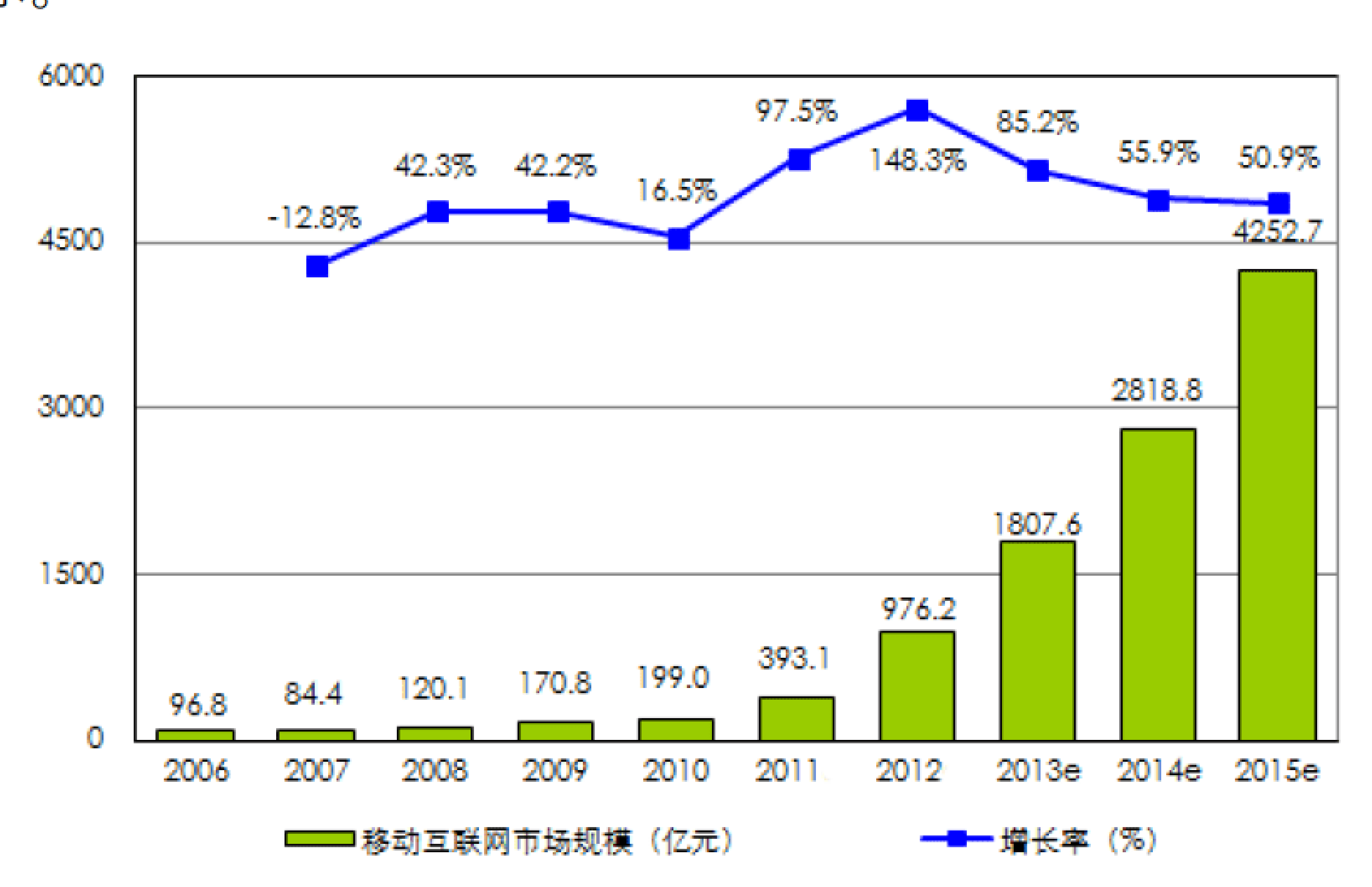


图 9-2　2006—2015 年中国移动互联网市场规模

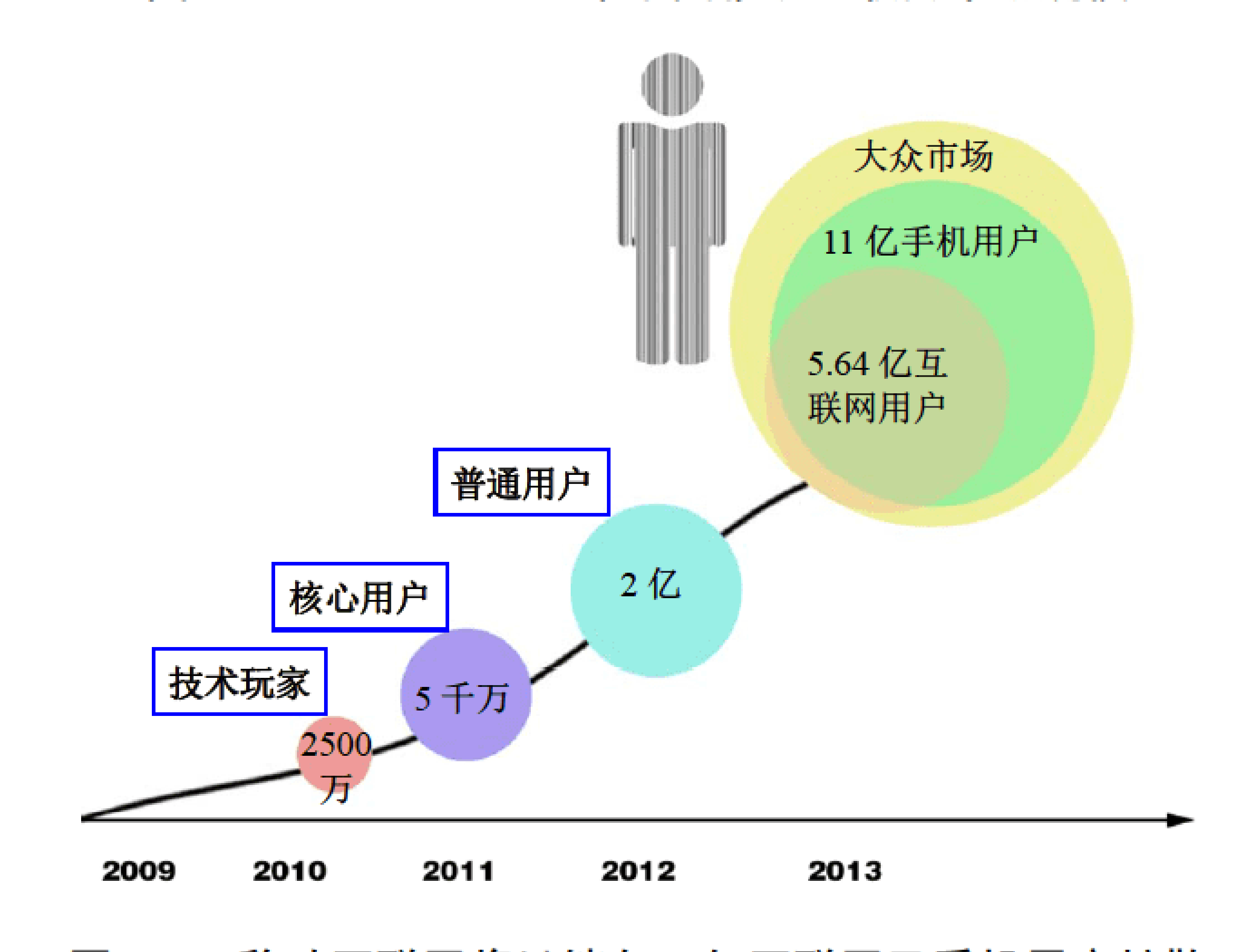


图 9-3　移动互联网将继续向 5 亿互联网及手机用户扩散

移动互联网正逐渐渗透到人们生活、工作的各个领域，短信、铃声下载、移动音乐、手机游戏、视频应用、手机支付、位置服务等丰富多彩的移动互联网应用迅猛发展，正在改变信息时代的社会生活。

我们尚无法确定万物是否皆数据，但是，在移动互联网时代，人类至少已经推开了这样一扇大门：通过对海量大数据的高效分析获得商业以及社会价值。大数据为移动互联网带来了新的价值，也为迈向物联网奠定了基础。

移动互联网成为大数据非常重要的来源，很多公司在移动互联网上面的产品，尤其是很多互联网公司，其产品数量都超过一半。例如，以微信、手机 QQ 为代表的即时通信类占到移动互联网总有效浏览时间的 18%，浏览器为 12%，在线视频为超过 10%，游戏为 11.65%。这些都是移动互联网快速被推广使用下形成的这样一些用户的应用平台，但这些用户应用平台也都是收集用户大数据新的来源。

在移动互联网的多 App 时代，大数据的“入口”概念是模糊的。每个用户都有其常用的若干个 App，并不断下载新的 App。在这样的情况下，谁控制了强大的后台，谁就能拥有强大的数据分析能力，从而推送或者显示精准信息。另外，手机的私人性和唯一性比电脑要更强，如果用户在多个 App 的行为能在后台被统一进行分析，自然可以更好地抽象出用户的特征和行为。

例如，你在京东商城或者亚马逊订了一件商品，那么机器就会将你的 ID 号码、送货地址、手机、电话、电子邮件以及收货时间等全部记录下来。如果你提交了物品评论，或者和好友在微博上进行了分享，同样也会被记录下来。

洞察这一切，就意味着梦寐以求的商机。移动互联网与社交网络的兴起将大数据带入新的征程，互联网营销将在行为分析的基础上向个性化时代过渡。创业公司用“大数据”告诉广告商什么是正确的时间，谁是正确的用户，什么是应该发表的正确内容等，这正好切中了广告商的需求。

9.2 互联网大数据应用案例

互联网是个变幻莫测的时代，抓住机遇才是王道，大数据的兴起让互联网企业找到了新的商机，将网站运营带入了精准营销时代。本节主要介绍互联网行业大数据的应用案例，希望对读者有一定的启发和学习价值。

9.2.1 【案例】大数据与互联网助力竞选总统

奥巴马胜选的原因不在于经济、外交政策或是妇女问题，而是赢在大数据。奥巴马借助超强的“大数据”能力成功连任，其背后几十人的数据分析与挖掘团队也浮出水面。

奥巴马的总统竞选运动也通过使用社交网络的各种数据功能完成了竞选，他们不仅通过社交网络寻找支持者，而且还通过社交网络召集了一批志愿军。

早在 2006 年，Facebook 就帮助总统候选人建立了个人主页，以便他们进行形象推广。2006 年 9 月，Facebook 全面开放，用户数量爆炸式增长，在年底达到 1200 万，这一过程恰好有利地推升了奥巴马的知名度。此后，奥巴马掀起了一系列的网络活动，在 Facebook、MySpace 等社交网站上发表公开演讲、推广施政理念，从而赢得大量网民支持，募集到 5 亿多美元的竞选经费。

奥巴马的数据分析团队建立了 4 条投票数据流，以了解关键州选民的详细情况。仅在俄亥俄州，数据分析团队就获得了约 2.9 万人的投票倾向数据。这是一个包含 1% 选民的巨大样本，这使他们可以准确了解每一类人群和每一个地区选民在任何时刻的态度。

2008 年，奥巴马赢在了互联网，当选为美国总统，被誉为首位“网络总统”。而 2012 年，奥巴马又赢在了大数据分析。如图 9-4 所示，美国总统候选人米特・罗姆尼与巴拉克・奥巴马展开第二次总统竞选辩论。

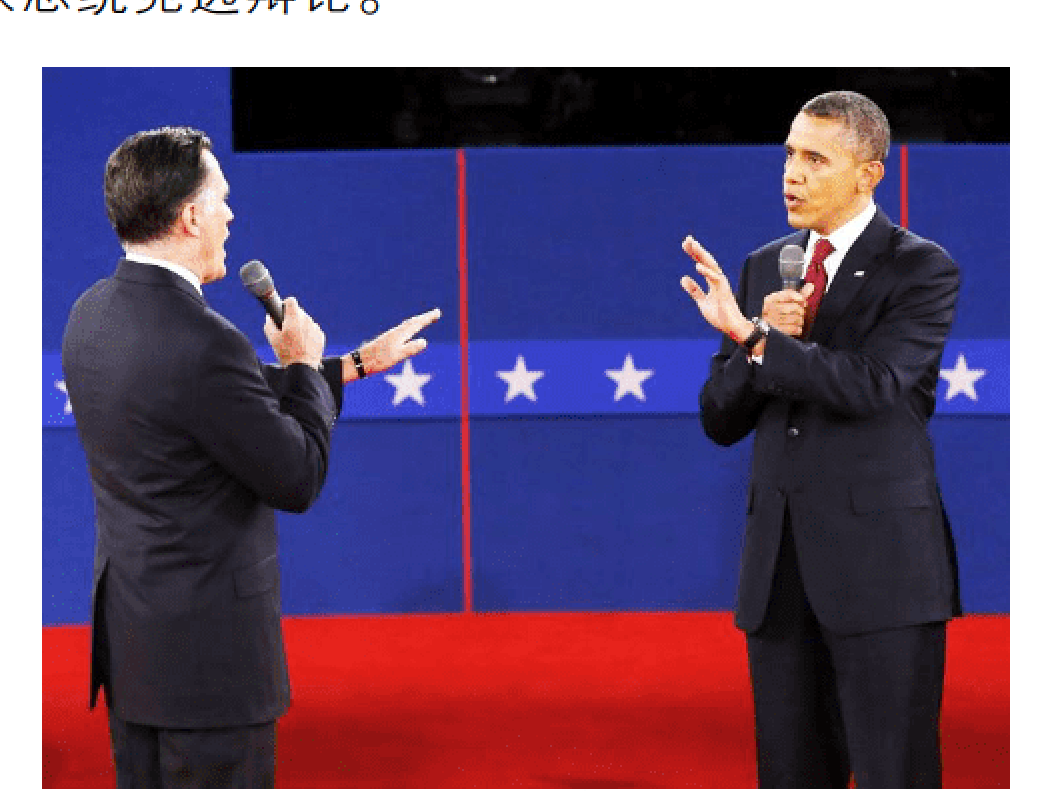


图 9-4　米特・罗姆尼（左）与巴拉克・奥巴马（右）展开总统竞选辩论

此次总统竞选，奥巴马的数据分析团队更动用了 5 倍于上届的人员规模，且进行了更大规模与深入的数据挖掘。这在帮助奥巴马获取有效选民、投放广告、募集资金方面起到了不可忽视的作用。数据分析团队分析来自各个途径的非结构化数据，包括网站、手机程序、志愿者和来自传统收集渠道的数据，他们能更全面地了解线上和线下的选民情况，准确地揣摩选民对各种话题的态度。

另外，掌握了数以 TB 的数据后，数据分析团队就能为选民建立更加准确的模型和计划。这意味着竞选活动将更有针对性，更多的网站注册人数、更多的电子邮件地址、更多的选票和献金。

数据分析团队不断试图挖掘选民的社交媒体信息，甚至还准备通过手机移动程序来改变传统的投票方式。通过定制手机程序的下载获取抽样用户，正在成为移动时代民意

测试员的新工作方式。随着数据科学家深入研究如何利用社交媒体数据提高预测准确性，在线民意分析的准确性无疑正随之提高，而其一旦与手机移动程序相结合，将对政治产生更为深刻的影响——候选人能对民意波动做出实时反应。

最终，“黑人平民”战胜了实力雄厚的对手，成为美国历史上第一位黑人总统，之后，在第二次的选举中更获得连任。此次选举被认为是美国民主的巨大进步，而互联网则提供了前所未有的实施手段，其中尤以 Facebook 为代表的社交网站最为突出，以至于有人将之戏称为“Facebook 之选”。

【案例解析】：从本案例可以看出，当“大数据”遇到“小数据”，大数据每次都会赢。数据驱动的决策对奥巴马——这位第 44 位总统的续任起到了巨大作用，这也是研究 2012 选举的一个关键元素。它也是一个信号——表明华盛顿那些基于直觉与经验决策的竞选人士的优势在急剧下降，取而代之的是数据分析专家与电脑程序员的工作，他们可以在大数据中获取洞察力。

9.2.2 【案例】Acxiom 用数据洞悉你的心理

现在越来越多的互联网公司在数据“矿山”中挖掘金矿，Acxiom 就是这群掘金者中的佼佼者。Acxiom 的主要业务是“基于数据的市场营销”，帮助企业精准定位它的潜在客户，将服务和产品卖给有需求的客户。

Acxiom 就是这样一个鲜为人知而又举足轻重的存在，它知道你是谁，它知道你住哪，也知道你喜欢什么，讨厌什么，事实上，在业内人口中，它有一个更为通俗易懂的名字——“数据精炼厂”。从种族、性别、体重、身高、婚姻状况、文化程度、政治倾向、消费习惯、家政开支到度假偏好，几乎每个美国成年人都能在 Acxiom 的数据全息图上找到自己的坐标。

Acxiom 可以利用一些信息来推测用户的生活方式、兴趣爱好和日常活动，例如，你的汽车品牌和使用时间、你的收入和投资状况、你的年龄、受教育程度以及邮政编码。除此之外，你最近离过婚吗，或者你刚刚变成了一名空巢老人？这些“人生大事”更可以将一个人从一个消费阶层转移到另一个，而这正是 Acxiom 及其广告客户的关键兴趣所在。Acxiom 称其可以通过分析数据来预测 3000 种不同的行为及心理倾向，比如说一个人会在某两个品牌间做出怎样的选择。

Acxiom 的大数据战略主要有 4 个方面，如表 9-2 所示。

表 9-2　Acxiom 的大数据战略

营销策略与分析	在现有的客户中找到收入增长的机会，识别并找到潜在的有价值客户
	发展洞察力，从而更有针对性地分配营销费用
	发现那些通过优化人员、流程、技术来降低成本的机会
	通过严格执行隐私政策来降低风险，保护客户免受欺诈

续表

多渠道营销	任意渠道的客户互动
	扩展和加强客户品牌意识的创意营销活动
	通过投资回报率指标量化营销效果
	符合现有最佳客户特征的新客户
精准定向营销	数据安全港：Acxiom 的隐私保护环境使广告商以及合作伙伴能够通过多媒体渠道准确地识别和屏蔽敏感信息
	精准定向渠道：通过与其他合作伙伴的合作，Acxiom 可以实现跨渠道传播高度协调一致的信息——不论是通过网络、手机还是电视等
	广告投放环境：帮助企业创造成熟的营销活动环境，在这样的环境中，企业的客户及潜在客户与企业选择的渠道及合作伙伴已经经过预匹配，这有助于企业进行有效的营销活动，增加营销信息的覆盖范围
	更准确的衡量：在客户定义细分层面上的所有响应渠道上，分析企业的客户及潜在客户对企业的营销活动的回应，通过在各种渠道跟踪销售转化数据来进一步优化企业的营销活动，帮助企业了解多种营销渠道的交叉影响
数据与数据库管理	建立数据库：Acxiom 的系统由经过市场检验过的标准组件构成。Acxiom 会对这些组件和系统进行个性化配置，满足企业的需求
	数据管理平台：使企业的营销活动覆盖更多的目标客户，提高企业的投资回报率
	营销活动管理：营销活动管理让营销者能够更精确、更有针对性地细分受众群体，以实现更个性化的互动
	IntegraLOOP 数据库营销解决方案：管理者希望所花费的营销投入能带来更大的市场回报。选择何种平台来管理数据库至关重要，明智正确的选择能帮助企业管理者更高效地进行客户数据管理、更便捷地进行操作、更全面地获得客户分析与决策支持。IntegraLOOP 数据库营销服务解决方案正是基于这些标准模块，再根据企业特有的业务需求加以客户化定制，包括业务规则定制、报表定制、业务系统集成、网站数据集成、客户服务系统集成等，为企业带来完美的数据库营销系统
	数据整合和质量：通过提高企业的数据库的搜索和识别功能，优化企业的数据库；还可以通过更精准的身份识别方案进行进一步优化
	数据优化：使用 InfoBase®立即了解特定客户的需求；使用 Personicx®进行数据优化并在各个市场上寻找客户

目前，Acxiom 正从微软、谷歌、亚马逊、MySpace 等 IT 业巨头“挖角”，旨在打造一个更强大、更多元的“消费行为预测复式平台”，通过对数据库的深耕细作，巩固其在投资者和客户当中的地位。Acxiom 的最大优势在于其过去 40 年中对“离线数据”的搜集和积累，这亦是它能够雄踞一方的秘诀所在。

【案例解析】：在本案例中，Acxiom 公司的解决方案有助于简化数据分析和管理，

并推动企业的营销计划。

但是，无论手法有多巧妙，这一切都是在客户本人毫不知情的前提下发生的。究其本质，这是“数据驱动时代”的不可承受之重。我们的生活“被挖掘、被提炼，然后被卖给出价最高的竞拍者”，蛰伏在暗处的数据巨兽在绕过当事人的情况下，与商家达成了某种“幕后交易”。

也许“大数据”时代的到来，会让每个人都陷入这样的困境，你的一举一动都被记录成数据，变为有价值的信息，但你又不可能离开这个世界，也难以离开媒介。

9.2.3 【案例】大数据为个性化用户体验撑腰

根据 2012 年的相关统计显示，在线视频已经超越社区交友和搜索服务跃升为互联网第一大应用。PPTV 聚力目前全平台月度活跃用户达 3.4 亿，每天的活跃用户超过 5000 万。目前，PPTV 聚力每天会产生数 10TB 包含用户行为数据、访问体验数据等在内的业务数据，针对在线视频业务运营的实际需要，这些大数据每天会被采集、汇总到一个分布式的技术平台上，再被应用到不同的业务领域之中。

对此，PPTV 聚力正努力超越数据解析，利用大数据与分析技术，改变思维定式，为用户提供真正个性化的服务体验。细心的老用户会发现，登录后均可看到“猜你喜欢”栏目，在这里，超过 35%的用户都能找到自己喜欢的视频，使你不会在浩瀚的视频节目里不知所措，而且缩短了视频搜索浏览时间，大幅提升了用户体验。

事实上，对于 PPTV 聚力带来的个性化视频推荐用户体验，大数据是功不可没的。目前，PPTV 聚力已经建成的数百台服务器规模的 Hadoop 集群是其大数据技术平台的核心。在其上运行着 Hive 开源数据仓库，基于 Storm 的分布式实时数据处理框架也已经开始部署。

对 PPTV 聚力来说，大数据的来源主要包括用户行为数据、工程技术数据以及后端的业务运营数据，如图 9-5 所示。

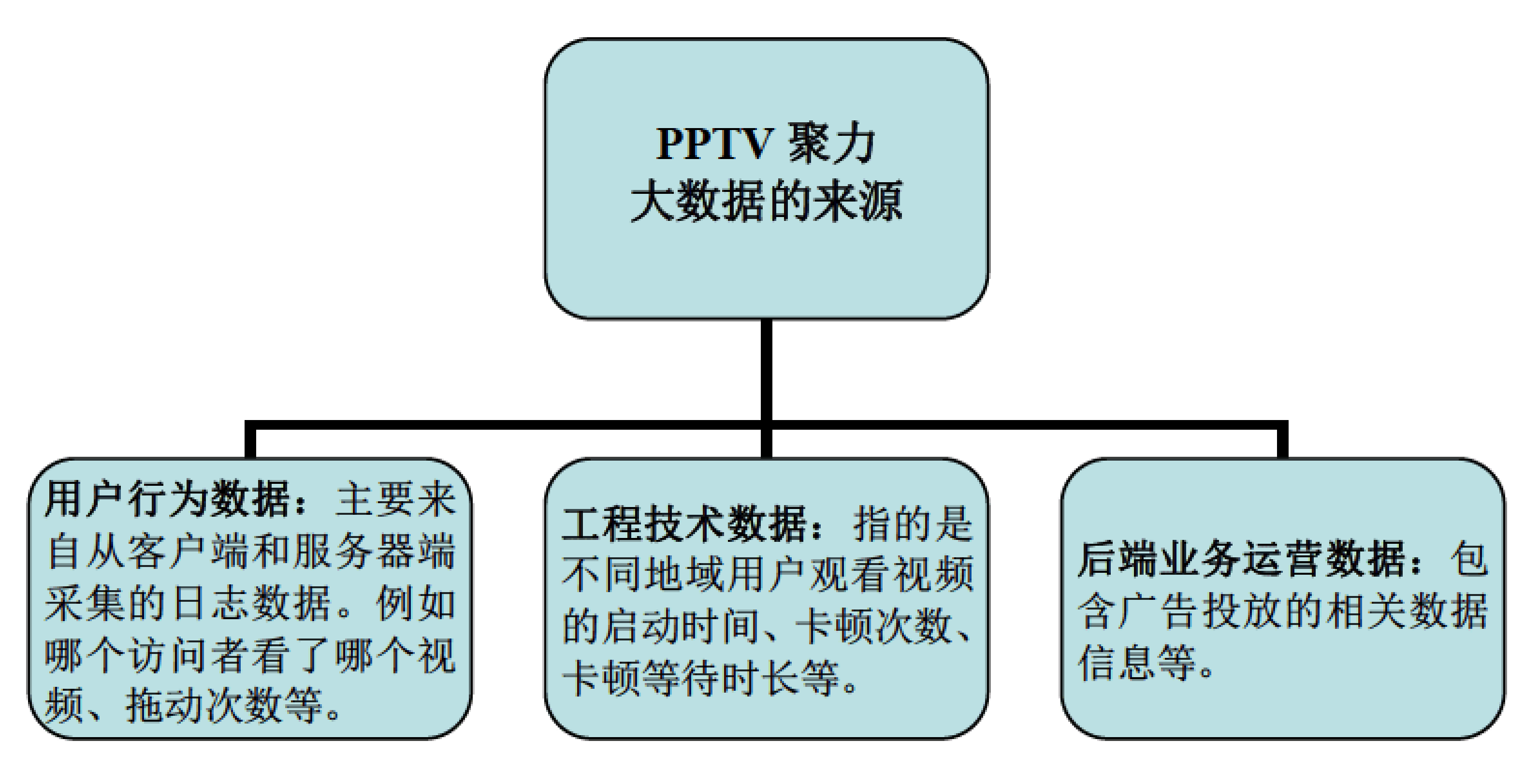


图 9-5 PPTV 聚力大数据的来源

这些数据组成了 PPTV 聚力丰富的大数据来源，而大数据的分析结果能直接应用于商业运营的调优。例如，我们购买了一部影视剧后，可以精确、实时地了解它在不同地区和时间段被观看的次数，以此优化后端的运营策略。另外，通过从不同的客户端所获取的访问连接数据，我们可以根据不同地区、不同时段的网络连接状况，用最低的成本向用户交付流畅的观看体验。

基于大数据技术平台，PPTV 聚力已经在广告的定向投放、频次控制等方面建立了相对成熟的策略和流程，并且注重在广告精准投放的同时，确保用户的观看体验。

【案例解析】: 在本案例中，通过对大数据的深入了解和熟练运用，PPTV 聚力更将与我们"如影随行"，打造一幅智能个性化用户体验全新蓝图，向着更优越的用户体验境界进发。

笔者认为，国内的视频网站，仍处于飞速发展阶段，可以考虑未来自建数据中心，提高数据处理能力，从网站的运营中发掘出更多信息，为用户提供更好的视频服务。

9.2.4 【案例】人人游戏网用大数据了解玩家

作为国内最大的网页游戏和智能手机游戏的研发、运营和发行商之一，人人游戏的大数据价值发现从结构化数据集起步，逐步向非结构化数据集延伸。成立于 2006 年的人人游戏坚持在"跨屏"技术创新领域的研发投入，同时也积极利用大数据技术优化整体业务运营。

近日，IBM 公司宣布正式与人人游戏在业务分析领域展开合作，通过部署全球领先的 IBM 商业智能和业务分析平台，利用创新大数据分析技术为人人游戏业务运营、企业管理、企业战略和企业文化注入全新动力。

人人游戏通过运用 IBM 的大数据解决方案，对企业内部数据的深刻分析和高价值运用，得以在互联网行业激烈竞争中脱颖而出，在高效应对多样化客户需求，提供针对性服务策略方面实现大步提升，真正实现了运营、管理"双创新"。

人人游戏的第一个动作就是上线"词云"应用。所谓"词云"，就是先对人人游戏玩家的在线聊天记录进行分词，汇总之后对玩家行为进行分析和展现。目前，"词云"已经在人人游戏的 4 款重点游戏中安家落户，随后有关玩家情绪的分析功能（通过关键词对应玩家的情绪指数）也将上线。

例如，"炸金矿"是人人游戏旗下"乱世天下"这款游戏中玩家参与度很高的一个玩法，玩家需要邀请一定数量的友人帮忙炸矿来赢取金币。但在节假日期间，这款游戏的参与度通常都会下降。数据分析人员通过"词云"应用分析后发现，节假日期间"求炸"成为玩家的聊天热词。数据分析人员也因此得知，并不是玩家不爱玩这个游戏，而是玩家在节假日邀请不到足够数量的友人帮忙炸矿。基于这样的分析，人人游戏可以在节假日期间对游戏规则进行调整。

在 2013 年，人人游戏已经基本上完成了基于 IBM Cognos 的 BI 系统整体建设。同时，其基于 Greenplum 社区版的分布式数据仓库也已初具规模。对人人游戏而言，这些都是获得 360° 用户视图的必要工作，而 360° 用户视图为其业务运营和决策所带来的价值则是实实在在的。

该平台的一大设计原则和优势是将报表分析平台与游戏业务模型（Acquisition Retention Monetization，ARM）紧密结合，通过 Cognos（如图 9-6 所示）强大的可视化报表和分析功能，以日、周、月的维度分析基于用户获取、存留和变现的海量数据，增进运营团队对于用户的了解，促进更有效的回访，及时调整运营的策略和推广重点。



图 9-6　Cognos 的分析界面

新的 BI 系统将人人游戏的业务模型更加清晰地呈现出来，对游戏业务覆盖用户获取、客户存留、客户付费的核心流程进行了优化，能够更准确地为业务决策提供参考。同时，BI 系统上线后，企业在开发和运维方面的投入也有所降低。财报显示，2013 年第一季度，人人在线游戏收入达 2670 万美元，同比增长 52.9%，占人人总营收的 57%。

【案例解析】：在本案例中，“词云”应用的上线是人人游戏对大数据的利用从结构化数据集向非结构化数据集延展的重要一步。

笔者觉得，大数据分析可能不会直接为网络游戏行业带来电商网站那样的可观收入，但其价值同样会体现在精准营销、客户体验优化等多个层面。当然，大数据团队所面临的最大挑战是数据的整合，把多来源的结构化、半结构化和非结构化数据整合在一起，很多企业还没有做到。另外，在企业内部和外部找到大数据的消费者，向他们营销大数据技术，同样是一项艰苦的工作。

专家提醒

将全面的大数据分析用在网络游戏中，能够有效提升玩家的留存率和转化指标，并且为游戏产品的研发提供指引。另外，个性化的精准营销同样与大数据分析紧密相关，像针对不同性别、不同年龄、不同地域人群的广告精准投放，背后都要依靠基于大数据的玩家特征分析。

9.2.5 【案例】迅雷用大数据抓“网络票房”

迅雷看看近日发布了迅雷看看电影院（付费频道）用户画像数据报告，报告中的一组突出数据是，90%以上影视 VIP 用户（付费会员用户）为男性，可见男女用户付费行为差异巨大，抓住了男性用户，就抓住了“网络票房”。

通过数据分析显示，影视 VIP 用户的主要需求是高清（高清/正版）、最新、大片/经典，其中高清占比最高，超过 80%，如图 9-7 所示。可见，高清画质已成为高端视频用户的一大重要需求。

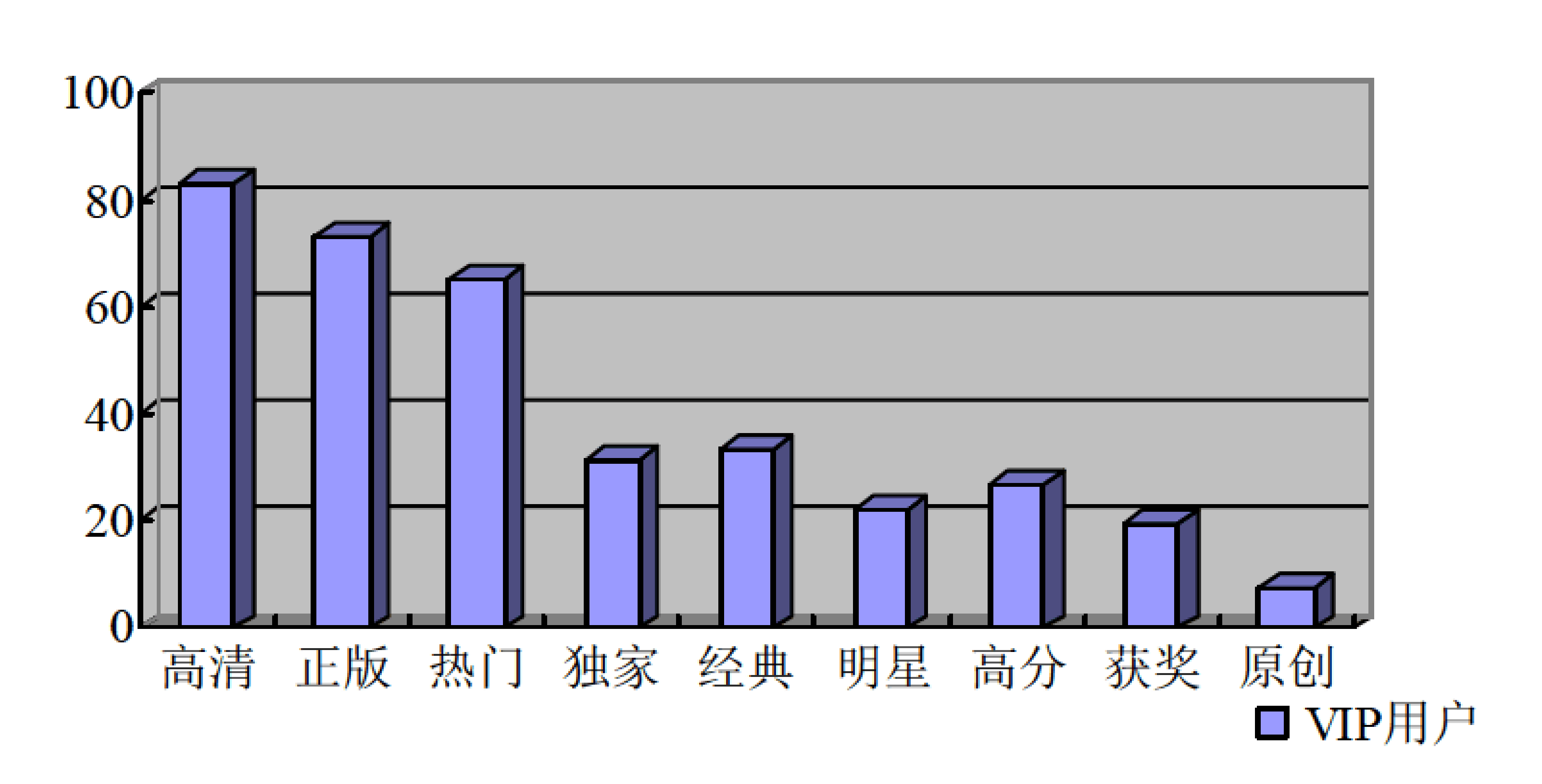


图 9-7　迅雷看看影视 VIP 用户主要需求

2013 年 12 月 20 日，迅雷集团 COO 黄艽首次代表集团对迅雷看看的战略规划作出明确部署。迅雷全线产品矩阵将大幅度提升对迅雷看看的支持，将实现 1.5 亿注册用户将与迅雷看看的大数据共享，迅雷看看将正式从流量平台向用户平台蜕变。

迅雷的数据分析模块目前有 500 多台服务器，4000 多个 CPU，存储 20PB 以上的数据，磁盘有 3000 多块，属于中型数据平台的规模。迅雷会把收集来的数据做成数据模型，最重要的数据模型是一个用户事件模型，所有的基于用户端的这种行为数据，都可以把它抽象成模型存进去。例如，每个产品的上线用户数，每个用户的活跃度，用户的黏性，以及某个产品的用户的地域分布，运营商分布等，活跃用户排名，最热的资源排名，如哪些电影最常被人看，发生某个事件所消耗时间是多少，或者某个商品的销售收入等。

另外，迅雷还构建了一个用户的“染色库”，用于记录几亿迅雷用户的特征、网络运营商类型、兴趣类标签、游戏类标签、影视类标签等。例如，游戏标签描述该用户喜欢什么类型游戏，影视类标签描述用户喜欢什么类型的影视。根据这些属性，迅雷可以更好地为用户服务。

除 iSO、Android 的手机与平板电脑端的产品之外，迅雷看看发布了电视端的 APP 产品——看看 TV。据悉，迅雷看看与 VIVO、AUDEX 等硬件厂商达成战略合作关系，力图在大数据领域开创一片新天地。

【案例解析】：本案例中的迅雷看看能否借助集团资源实现用户平台的转型，或将影响 2014 年网络视频行业格局。由此可见，随着移动互联网战役的拉开，移动端多屏的流量之争已经剑拔弩张。

笔者觉得，其实用户才是互联网真正的价值所在，利用大数据来挖掘用户属性和行为的视频互动营销，才是网络视频最深刻最有效的营销。

9.2.6 【案例】腾讯用微信展开大数据“首战”

微信是腾讯目前最成功的移动互联网应用，也是互联网历史上增长最快的新软件。如果 QQ 和 Qzone 是腾讯 PC 端的大数据开放平台，那么微信将成为腾讯移动端的大数据开放平台。

就拿笔者自己来说，我会用微信跟好友和同事联系，看下几个群里大家在讨论些什么，再刷刷朋友圈看看大家分享了些什么好东西，每天花在微信上的时间累计起来至少超过两小时以上。可以说这些事情基本是目前每个微信用户都在做的，至多是因为圈子或兴趣爱好等不同看到的内容不一样，但是这些信息基本上完整地描述了我一天的行为，同时还带着地理位置。

腾讯拥有最多的社交大数据，前期的思路是用数据分析改善自有产品，注重 Qzone、微信、电商等产品的后端数据打通。腾讯云移动分析平台已接入了微博、QQ 游戏、QQ 互联、空间、手机 QQ 多个平台的数据，现在另外一块相对封闭但是极具价值的微信数据也被打通了。

腾讯的大数据价值如何释放，如何变现？笔者认为，最优的途径是将数据分析成果共享给开发者，让开发者二次挖掘，腾讯则获得对应的收益。具体的方式有很多种，例如按照特权接口收费，按照接口调用次数收费，按照定制化功能收费。被阿里巴巴收购的友盟、AWS、围绕微博的一些数据分析公司做的也是类似的事情。

2013 年 8 月，微信公众平台增加了一项新功能——数据统计功能，包括用户分析、图文分析、消息分析和开发支持 4 个模块。

（1）用户分析。管理者可以在这个模块了解到账号的用户增长情况及用户属性，如图 9-8 所示。用户增长关键指标包括新增人数、取消关注人数、净增人数、累计关注

人数等，以相应的曲线图和数据表来显示数量发展趋势。在用户属性中，可以看到用户的性别、语言、省份分布数量以及各自所占的比例。



图 9-8　用户分析功能界面

（2）图文分析。包括图文群发和图文统计两部分。管理者可以在此看到图文消息中的每篇文章有多少用户接收、图文页阅读数量、原文页阅读次数以及文章的分享转发人数和次数等。此外，后台也提供了按照图文页阅读人数、分享转发人数进行排序的功能，这样一来，相应的时间段内，哪些文章最受欢迎一目了然。

（3）消息分析。这里主要是查看用户向公共账号发送的消息数统计，由此管理者可了解读者与账号的互动情况。

（4）开发支持。使用开发模式的管理者可以在此查看接口调用的相关统计，例如调用次数、失败率和平均耗时等。

【案例解析】：在本案例中，通过微信公众平台的数据统计功能，可以轻松掌握公共账号的实际运营效果，这对公众账号管理者来说无疑是一个好消息。

在这个大数据爆发的时代，每个人的行为规律都被记录成数据，对这些数据都可以找到规律并做出分析。不可否认，微信通讯录已经慢慢等同于笔者的手机通讯录，里面也不再仅仅是好友和家人，还有同事、客户等社会关系在里面，另外还有微信群、公众账号等，如何管理、分享或者搜索有赖于开发者的智慧。

10

零售：打响大数据之战

学前提示

俗话说：“他山之石，可以攻玉。”大数据里面包含了企业运营的各种信息，如果能对它们进行及时有效的整理和分析，就可以很好地帮助企业进行经营决策，为企业带来巨大的增值效益。零售企业要学会利用自己手中的海量数据，推动企业的发展。

要点展示

- 零售行业大数据解决方案
- 零售行业大数据应用案例

10.1 零售行业大数据解决方案

当你惊叹于淘宝通过对以往消费的记录，准确推送你所需的小众商品的时候，恭喜你已经感受到大数据时代的来临。在大数据时代，我们在网络上的任何一次点击都可以被完整地记录和保存，而零售企业则通过对这些数据的高效分析，准确预判我们的消费行为、消费心理等，并推送相应的产品或服务。而实际上，目前多数大数据并未被采集到，即使采集到，其价值的开发也远远不足。

10.1.1 大数据对零售行业的影响

近年来，互联网技术改变着各行各业，零售行业自然难逃厄运。随着电子商务不断发展，消费者的购物习惯悄然生变。在中国，零售商、制造商、个体户等均可在淘宝网、京东商城这类第三方平台开展电子商务业务，因此，消费者也有了更多选择和主动性，这给传统零售产业带来巨大的冲击。

安吉尔知识网络公司（Edgell Knowledge Network）是一家调研及内容服务公司，其在 2012 年 5 月至 6 月对北美零售经理进行了一项调查，具体如图 10-1 所示。结果显示，只有 17%的零售经理不知道“大数据”概念；其余的受访者对“大数据”具有不同程度的熟悉，有 10%的人说自己理解“大数据”的理念，但不确定此概念如何对零售产生影响。

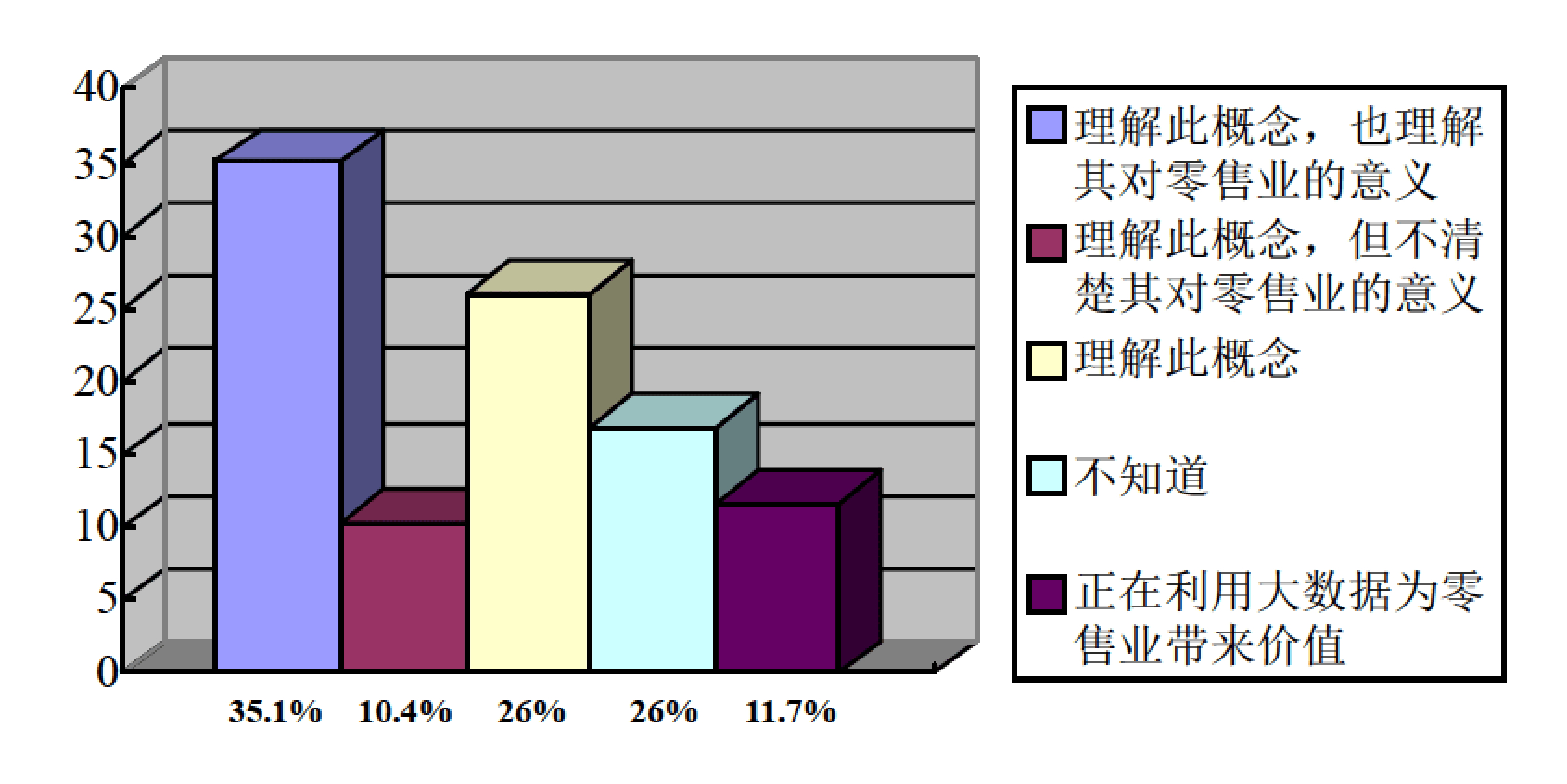


图 10-1 北美零售经理对大数据的了解程度

在大数据时代，智能零售可以分为四等份，分别是客户数据资源、社会数据资源、市场数据资源以及供应数据资源。智能零售能够生产出源源不断的数据，创造出数百万的交易以及数以亿计的交互。大数据及分析环境中的投资收益将通过传统客户忠诚度、收益增长、成本削减以及新业务模式而货币化。

10.1.2 大数据对零售行业的挑战

随着中国大型连锁零售企业开始规模化经营和跨区域发展，“用 IT 去做零售业”已经逐渐成为零售业的重要经营理念之一。

零售商在处理大量数据方面已经有很长的历史了，多年来条形码和库存管理任务都需要信息分析，但是“大数据”对那些认为自己拥有良好数据分析能力的零售商也提出了挑战。

近年来，我国的零售业正处在成长与巨变的风口浪尖，呈现出如下发展趋势：零售变革速度加快，市场空间饱和新旧产业形态并存，外资企业长驱直入，企业经营日趋同质化，盈利模式单一等。零售企业迫切需要提高自身的核心竞争力，其主要策略是外拓和“内敛”。

- 外拓：主要是指通过并购和自营店面数量的扩张实现规模化发展。
- “内敛”：主要是指通过加强 IT 信息化建设来实现内涵式增长。

Edgell Knowledge Network 通过调查发现，46%的零售商认为处理大量数据是其最大的挑战，而 34%的零售商表示仅仅大量的数据类型就占据了自己很多的注意力，20%的零售商认为数据产生过于频繁，对自己来说是个麻烦，如图 10-2 所示。

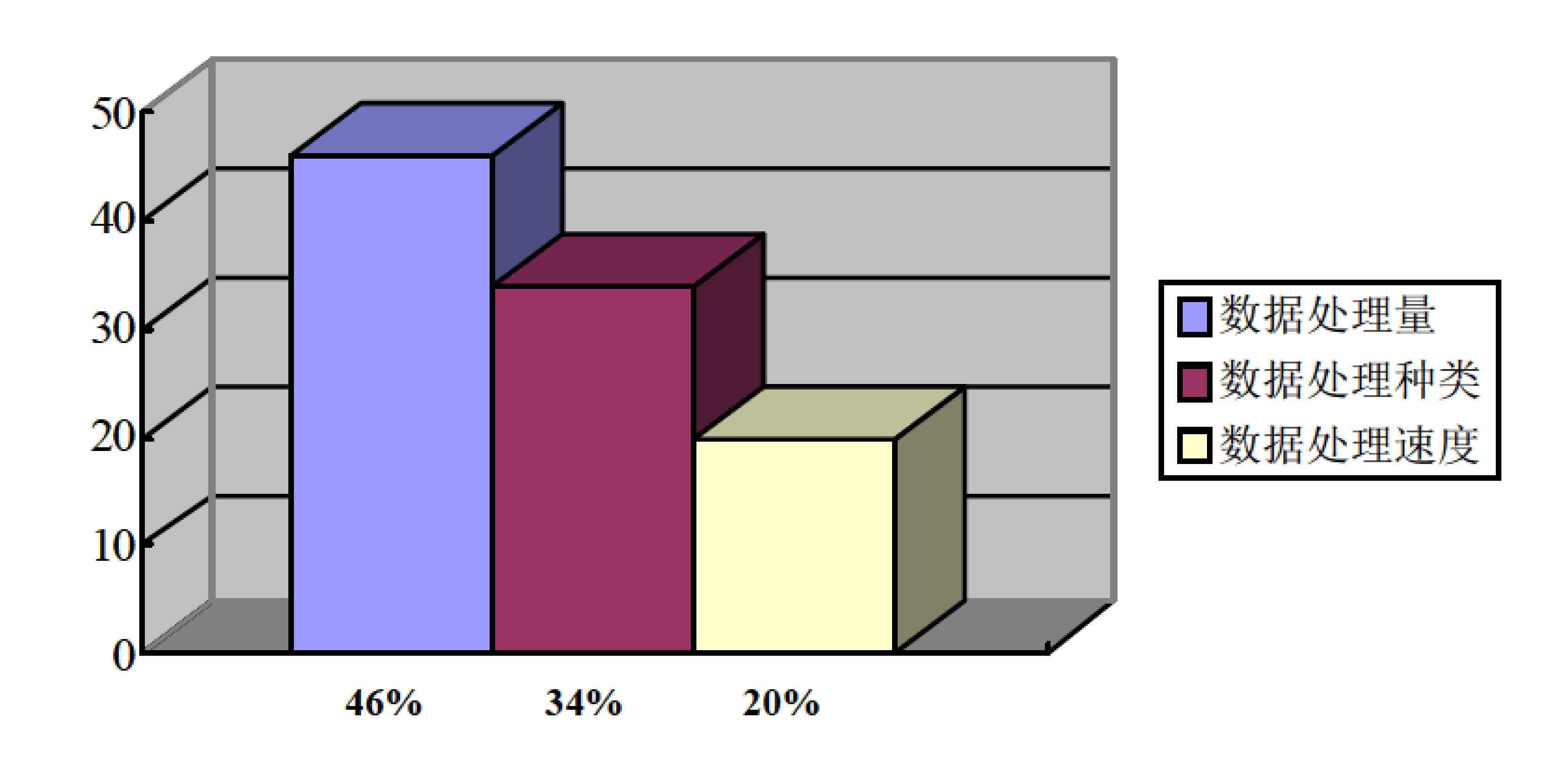


图 10-2 北美零售商认为管理“大数据”带来的最大挑战

如何培养忠实的消费群，并充分挖掘客户信息中所蕴藏的商业价值，如何用数据为企业的经营提出实时的决策指导，已经成为零售企业长足发展的迫切需求，也是零售企业面临的挑战。

专 家 提 醒

笔者认为，阻碍零售商把更多的资源投入“大数据”领域的因素应该是潜在的收益和投资回报仍然不明朗。

10.1.3 大数据对零售行业的价值

如今，中国零售业面临着巨大的挑战和困难，整个行业都在积极探寻发展出路。此时，一个新的关键词出现了，让整个行业看到了新的曙光，它就是“大数据”。

毫无疑问，我们已经进入了大数据时代，面对海量、碎片化的数据，零售企业该怎么利用和管理，为企业的发展提供帮助，可能是一些管理者正在思考的问题。笔者认为，大数据对零售行业的价值主要体现在 6 个方面，如图 10-3 所示。

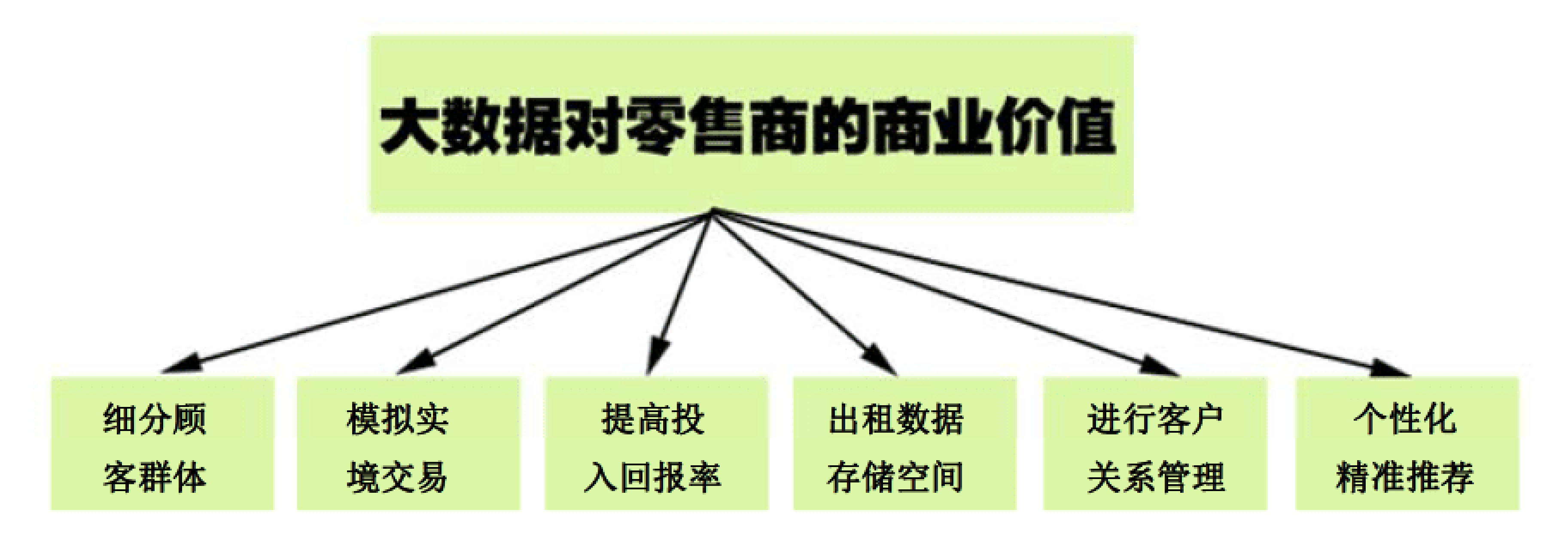


图 10-3 大数据对零售行业的价值体现

专 家 提 醒

笔者认为，个性化精准推荐是零售商运用大数据的最重要“法宝”。以日常的“垃圾短信”为例，信息并不都是“垃圾”，因为收到的人并不需要而被视为垃圾。通过用户行为数据进行分析后，可以给需要的人发送需要的信息，这样“垃圾短信”就成了有价值的信息。在日本的麦当劳，用户在手机上下载优惠券，再去餐厅用运营商的手机钱包优惠支付。运营商和麦当劳搜集相关消费信息，例如经常买什么汉堡，去哪个店消费，消费频次多少，然后精准推送优惠券给用户。

大数据对零售企业的最大价值是，将零售策略与大数据技术进行结合，最大程度地编制前瞻性的零售策略，确保销售计划的实现。因此，零售企业可以根据大数据的特性，主动地在业务数据产生的同时做出相应的策略应对，为企业赢得更多的时间和市场策略调整空间。要做到这一点，零售企业的需要注意以下 4 个方面：

（1）转换态度。企业的领导者首先要重视大数据的发展，重视企业的数据中心，把收集顾客数据作为企业营销的第一目标。

（2）做好准备。对企业内部人员进行培训及建立收集数据的软硬件机制。

（3）制定原则。以业务需求为准则，确定哪些数据是需要收集的。

（4）规划目标。确认在企业已有的数据基础上或者未来方向前提下，如何达成前三项目标的基础建设方案。

目前，一些 IT 软件开发运营商也已经针对传统零售企业推出了云服务的基础平台，

为中小微型商业企业提供了大型企业和超大型企业同样的基础环境及系统架构，小的零售企业只需清晰地规划出自己的目标和适合的步骤，使用云平台按需付费即可，大可不必进行巨大的初始投入。

也许在不久的将来，你可以感受这样一个场景：你和家人在家中正在列出自己出去购物的清单，一家商场的客服会“恰到好处地”发来短讯，提醒你新到了一些货品，而这些货品很可能“恰好”也在打折，而这些商品也“恰好”正是你想购买的商品，甚至连你没有想到而需要购买的商品，都在通知的清单中。笔者认为，这或许是对大数据这门“内功”应用到炉火纯青的地步的表现。

在大数据时代，一切似乎都变得数据化，如何利用这样大量的数据做到以顾客需求为上，就有待各个零售企业“八仙过海各显神通”了。零售业用好大数据，可以焕发新的生机，进入蓬勃发展的新时期。

10.2 零售行业大数据应用案例

值得关注的是，当国内的大数据研究还停留在概念阶段和初步应用阶段时，国外的一些企业已经在如火如荼地运用大数据，并带来了可观的经济效益。本节主要介绍零售行业大数据的应用案例，希望对读者有一定的启发和学习价值。

10.2.1 【案例】ZARA：可以预见未来的时尚圈

ZARA 是西班牙 Inditex 集团旗下的一个子公司，它既是服装品牌，也是专营 ZARA 品牌服装的连锁零售品牌，为全球排名第三、西班牙排名第一的服装商，在世界各地 56 个国家，设立了超过两千多家的服装连锁店，如图 10-4 所示。



图 10-4 ZARA 服装连锁店

走进 ZARA 的店内，可以发现柜台和店内各角落都装有摄影机，店经理随身带着

PDA（Personal Digital Assistant，又称为掌上电脑）。当消费者向店员反映："这个衣领图案很漂亮"、"我不喜欢口袋的拉链"这些细微末节的细项时，店员都会向分店经理汇报。经理通过 ZARA 内部全球资讯网络，每天至少两次给总部设计人员传递资讯，由总部作出决策后立刻传送到生产线，改变产品样式。

每天关店后，销售人员都会盘点货品上下架情况，并对客人购买与退货率做出统计，再结合柜台现金资料和交易系统做出当日成交分析报告，分析当日产品热销排名，然后数据会直接传送至 ZARA 的仓储系统。

ZARA 为了增加网络巨量资料的串连性，2010 年在 6 个欧洲国家成立网络商店，并于 2011 年又分别在美国、日本推出网络平台，除了增加营收，线上商店强化了双向搜寻引擎、资料分析的功能。

ZARA 通常先在网络上举办消费者意见调查，再从网络回馈中，撷取顾客意见，以此改善实际出货的产品。ZARA 的网络平台不仅会回收意见给生产端，让决策者精准找出目标市场；也对消费者提供更准确的时尚讯息，双方都能享受大数据带来的好处。同时，网络商店还为 ZARA 至少提升了 10%的营收。

通常情况下，会在网络上搜寻时尚资讯的人，对服饰的喜好、资讯的掌握、催生潮流的能力，比一般大众更胜一筹。ZARA 也紧紧掌握了这一群人的动态信息，将网络上的海量资料看作实体店面的测试指标。再者，会在网络上抢先得知 ZARA 资讯的消费者，进实体店面消费的比率也很高。

ZARA 推行的海量资料整合，后来被 ZARA 所属英德斯集团底下 8 个品牌学习应用。可以预见未来的时尚圈，除了台面上的设计能力，台面下的"资讯/数据大战"将成为更重要的"隐形战场"。

运用大数据分析，ZARA 最短 3 天可以推出一件新品，一年可推出 12000 款时装。ZARA 平均每件服装价格只有 LVHM 的四分之一，但是，回看两家公司的财务年报，ZARA 税前毛利率比 LVHM 集团还高 23.6%。

【案例解析】：在本案例中，ZARA 通过收集海量的消费者意见，做出生产销售决策，这样的做法大大降低了存货率。同时，根据这些电话和电脑数据，ZARA 可以分析出相似的"区域流行"，在颜色、版型的生产中，做出最靠近客户需求的市场区隔。

专家提醒

市场区隔（Market Segment）是将消费者依不同的需求、特征区分成若干个不同的群体，而形成各个不同的消费群。市场区隔不仅是静态的概念，也是动态的过程。它是了解某一群特定消费者的特定需求，通过新产品或新服务或新的沟通形式，使消费者从认知到使用产品或服务并回馈相关信息的过程。

"大数据"最重要的功能是缩短生产时间，让生产端依照顾客意见，于第一时间迅速修正。"大数据"运营成功的关键，是资讯系统能与决策流程紧密结合，迅速对消费

者的需求作出回应和修正，并且立刻执行决策。

10.2.2 【案例】沃尔玛：大数据帮你选好购物单

前面的章节已经讲了沃尔玛的数据中心基础构建，下面就来分析一下沃尔玛是如何利用大数据来助力零售业务的。50 年前，山姆·沃尔顿在阿肯色州的罗杰斯开创了第一个沃尔玛折扣商店，如今这家折扣零售商已经成为跨国公司。

下面列出了 18 个关于沃尔玛的事实。

事实 1：2012 年沃尔玛的销售额达 4440 亿美元，这个数字比奥地利的 GDP 多 200 亿美元。如果沃尔玛是一个国家的话，它将是第 26 个世界最大的经济体。

事实 2：沃尔玛有全球雇员 220 万，相当于休斯敦人口，仅在美国就雇用了 140 万员工。

事实 3：如果把沃尔玛比作一个军队，它将是仅次于中国的世界第二大军队。

事实 4：沃尔玛相当于家得宝、克罗格、塔吉特、希尔斯、好食多和凯马特这些企业的组合。

事实 5：平均每个 4 口之家每年在沃尔玛花费超过 4000 美元。

事实 6：沃尔玛有分布在 27 个国家的 10400 家商店，每周的顾客超过两亿。

事实 7：美国人花在食品杂货上的每 4 美元中，就有 1 美元是花在沃尔玛。

事实 8：2012 年，首席执行官迈克尔·杜克年薪是 3500 万美元，每小时的工资比一个全职雇员全年赚的还多。

事实 9：2009 年，沃尔玛销售最多的商品是香蕉。

事实 10：2001 年—2006 年，中国对沃尔玛的出口占美国对华贸易逆差增长（growth）的 11%。

事实 11：将沃尔玛的所有零售商店空间平摊在同一个地方，将超过 9 亿平方英尺，达到 34 平方英里，大约是曼哈顿的 1.5 倍。

事实 12：沃尔玛的停车场占地规模相当于佛罗里达州的坦帕市。

事实 13：2000 年，沃尔玛起诉是 4851 次，相当于每两小时一次。

事实 14：90%的美国人生活中，15 英里范围内就有一个沃尔玛店。

事实 15：沃尔玛家族把 2%的收入捐给了慈善机构。比尔·盖茨捐了 48%的净资产，而沃伦·巴菲特捐了净资产的 78%。

事实 16：每 10 万居民中新增加一个沃尔玛巨型商场，就使这些居民的平均体重指数增加 0.25 个单位，肥胖率增加 2.4%。

事实 17：全球卫星定位系统装置 Telenav 中，最常见的输入目的地是沃尔玛。

事实 18：沃尔玛有大约 4700 个（90%）国际商店不使用沃尔玛的字号，包括墨西哥的 Walmex、英国的阿斯达、日本的西友、印度的 Best Price。

从以上数据可以看出，沃尔玛本身就是一个庞大的数据库，可以用于商业上的各种分析和应用。

2011 年 4 月，沃尔玛以 3 亿美元高价收购了一家长于分类的社群网站 Kosmix。Kosmix 不仅能收集、分析网络上的海量资料（大数据），并且结合沃尔玛商场顾客的结账资料等数据，它还能将这些资讯个人化，提供采购建议给终端消费者。这意味着沃尔玛使用的大数据模式，已经从“挖掘”顾客需求进展到能够“创造”消费需求。

沃尔玛利用 Kosmix 打造了一套完整的零售大数据系统——“社交基因组（Social Genome）”，它还可以连接到 Twitter、Facebook 等社交媒体。数据工程师从每天热门消息中，推出与社会时事呼应的商品，创造消费需求。分类范围包含消费者、新闻事件、产品、地区、组织和新闻议题等。值得注意的是，如果沃尔玛能够通过社交网络的大数据，掌握消费者行为，或许它能重新定义消费的方式。

为了得到便利和快捷的支付体验，沃尔玛推出了可以让消费者进行智能手机支付的应用软件 Walmart App，如图 10-5 所示。沃尔玛通过对用户过去购买数据的分析，在用户打开 Walmart App 之后就能自动生成用户的购物单，预判他们想买的商品。



图 10-5　Walmart App

目前，Walmart App 已经含有购物单的功能，能告诉顾客他们想要货品的位置，而且还发放类似商品的电子优惠券。沃尔玛还在测试一款名为“Scan and Go”的系统，用户只要在手机上挨个扫描商品，然后在收银台扫一下手机就可以买单走人了，再也不用排长长的队了。

沃尔玛全球移动部门的掌门人 Thomas 表示：“完美的购物单就是你根本不用动手，你一打开它就在那里了，这就是我们想要的。”

专家提醒

沃尔玛在对消费者购物行为进行分析时发现，男性顾客在购买婴儿尿片时，常常会顺便搭配几瓶啤酒来犒劳自己，于是推出了将啤酒和尿布捆绑销售的促销手段。如今，这一“啤酒 + 尿布”的数据分析成果也成了大数据技术应用的经典案例。

【案例解析】：在本案例中，沃尔玛结合社交网络媒体和移动 APP，也是为了进一步提高其对大数据的分析、应用能力，将其对大数据的应用能力提升到一个全新的境界。

零售商对个人消费数据进行分析，用于预测“一系列高度敏感的个人属性”，包括性倾向、种族、宗教和政治观点、健康状况、饮食习惯、性格特征、怀孕状况、休闲娱乐追求、父母离异、年龄和性别等。笔者认为，零售商同时还要注意大数据可能带来的风险。例如，从本质上讲，像沃尔玛这样的公司会越来越多地使用数据，包括真实和预测数据，从而将人群进行分类，一些低收入阶层类别遭受较差待遇的风险在增加。

10.2.3 【案例】淘宝：开放“数据魔方”的秘密

2010 年 3 月，淘宝开放网站所有的交易数据，并将这一计划命名为“数据魔方”。商家、企业及消费者将可以分享到其海量原始数据，数据开放将有原则、分层次地进行。淘宝还将与第三方专业研究机构合作，为商家带来基于数据之上的分析、解读、业务建设等服务，协助商家培养其通过读数据指导业务的能力。

据悉，每天有数以万计的交易在淘宝上进行，与此同时相应的交易时间、商品价格、购买数量会被记录，更重要的是，这些信息可以与买方和卖方的年龄、性别、地址甚至兴趣爱好等个人特征信息相匹配。各大中小城市的百货大楼做不到这一点，大大小小的超市做不到这一点，而互联网时代的淘宝却可以轻易做到。

淘宝数据魔方就是淘宝平台上的大数据应用方案。通过这一服务，商家可以了解淘宝平台上的行业宏观情况、自己品牌的市场状况、消费者行为情况等，并可以据此进行生产和库存决策，而与此同时，更多的消费者也能以更优惠的价格买到更心仪的宝贝，如图 10-6 所示。



序号	关键词	搜索人气	搜索指数	占比	点击指数	商城点击占比	点击率	当前宝贝数	转化率	直通车
1	婚纱	205,280	864,611	44.58%	607,317	49.46%	69.74%	1,160,079	0.15%	1.74
2	婚纱 拖尾	24,365	75,423	3.7%	71,580	29.93%	94.8%	90,874	0.17%	1.25
3	婚纱 新款	21,901	80,349	3.95%	56,327	50.31%	69.6%	175,351	0.15%	1.35
4	婚纱 抹胸	19,899	23,920	1.15%	38,278	15.78%	161.56%	147,939	0.19%	1.31
5	婚纱 一字肩	14,310	27,818	1.34%	33,902	29.48%	122.36%	20,923	0.25%	1.04
6	婚纱 礼服	11,128	31,032	1.49%	16,137	43.75%	51.31%	427,002	0.04%	1.32
7	婚纱 v领	9,451	11,662	0.55%	13,333	14.33%	114.64%	11,741	0.1%	0.81
8	婚纱 公主	8,563	25,645	1.23%	16,102	44.59%	62.19%	151,943	0.1%	1.29

图 10-6　淘宝数据魔方界面

今年春天流行穿什么？喝什么？玩什么？网上最热销的品牌，最热搜的关键词又是什么？其实，以上问题都能通过淘宝数据魔方来一一解答。

淘宝网利用大数据统计分析得到了有趣的结果，当然这些分析更为卖家勾画出了他们潜在的客户类型图，从而实施精准的市场营销战略。例如，从在淘宝指数中查询“花露水”的结果可知，如果消费者决定在淘宝上购买花露水，他很有可能会购买驱蚊液、痱子粉，而很少去考虑其他驱蚊产品。

类似信息有多种用途，例如商家扩大或缩小经营范围时，可以藉此来选择扩大或缩小商品的类别；搞促销活动时，商城运营人员可以藉此选择促销的范围乃至不同商品的促销力度等。

多数卖家会先把店铺运营目标放在“卖货”上，之后才是“做品牌”。但即便是初级阶段的“卖货”目标，也要做好定位。淘宝数据魔方作为行业数据工具，主要的作用就在于“行业定位”。

其实，互联网的竞争就是圈住用户能力的竞争，淘宝依靠开放数据平台策略，让更多的人聚集到淘宝，使用他的服务，这首先就是人气上的胜利。此外，随着淘宝用户群的壮大，各种增值服务应运而生，而且淘宝也已经进行了有效的布局，包括阿里软件、阿里妈妈都是为此做的布局。

【案例解析】：在本案例中，随着淘宝用户数量的不断攀升，交易量的不断增加，淘宝必须要升级数据中心，增加数据中心的处理能力，从而提升网友购物体验。这就和沃尔玛、家乐福，为提高用户购物效率，减少付款排队等待，增加付款台，提升金融系统处理速度是一样的道理。

在笔者看来，淘宝数据魔方中也蕴含了电子商务行业的业务流程，每个维度都是站在店主关注的角度来设计，而且还帮助店主了解行业状况、目标群体、年龄结构、性别构成、上网时间、购买时间等，剩下的就是店主如何用数据来挖掘商机了。这个过程就好像做一道菜，淘宝数据魔方提供了大量的新鲜蔬菜和佐料，而且帮助用户做好一切下锅的准备。

专 家 提 醒

阿里信用贷款是，阿里巴巴通过掌握的企业交易数据，借助大数据技术自动分析判定是否给予企业贷款，全程不会出现人工干预。据悉，截至目前阿里巴巴已经放贷300多亿元，坏账率约0.3%左右，大大低于商业银行。

10.2.4 【案例】Target：准确判断哪位顾客怀孕

美国的出生记录是公开的，等孩子出生了，新生儿母亲就会被铺天盖地的产品优惠广告包围。因此，孕妇对于零售商来说是个含金量很高的顾客群体，但是她们一般会去

专门的孕妇商店购买孕期用品。

如果 Target 能够赶在所有零售商之前知道哪位顾客怀孕了，市场营销部门就可以早早地给他们发出量身定制的孕妇优惠广告，早早圈定宝贵的顾客资源。为此，Target 的市场营销人员求助于 Target 的顾客数据分析部要求建立一个模型，在孕妇第 2 个妊娠期就把她们给确认出来。可是怀孕是很私密的信息，如何能够准确地判断哪位顾客怀孕了呢？

不久后，Target 市场营销部经理 Andrew Pole 从公司的一个迎婴聚会（baby shower）上找到了“入口”。原来，迎婴聚会通过一个登记表记录了顾客的消费数据。Andrew Pole 从 Target 商品数据库的数万类商品和存放交易记录的数据仓库中挖掘出 25 项与怀孕高度相关的商品，制作“怀孕预测”指数，并以此可以推算出预产期，抢先一步将与孕妇相关的产品推送给客户。

为了不让顾客觉得商家侵犯了自己的隐私，Target 把孕妇用品的优惠广告夹杂在其他一大堆与怀孕不相关的商品优惠广告当中。

下面看一个关于 Target 的真实故事：美国一名男子闯入他家附近的一家 Target 连锁超市，并对店员抗议道：“你们竟然给我 17 岁的女儿发婴儿尿片和童车的优惠券。”店铺经理立刻向来者承认错误，但是其实该经理并不知道这一行为是总公司运行数据挖掘的结果。一个月后，这位父亲来道歉，因为这时他才知道他的女儿的确怀孕了。Target 比这位父亲足足早了一个月知道他女儿怀孕的情况。

根据这个“大数据”模型，Target 制订了全新的广告营销方案，结果 Target 的孕期用品销售呈现了爆炸性的增长。Target 的“大数据”分析技术从孕妇这个细分顾客群开始向其他各种细分客户群推广，从 Target 使用“大数据”的 2002—2010 年间，Target 的销售额从 440 亿美元增长到了 670 亿美元。

【案例解析】：在本案例中，Target 是基于数据挖掘所做的用户行为分析的结果。如果不是在拥有海量的用户交易数据基础上实施数据挖掘，Target 不可能做到如此精准的营销。然而，正是因为对于数据挖掘的充分应用，Target 才能在低迷的美国经济环境下持续发展。

可以想象的是，许多孕妇在浑然不觉的情况下成了 Target 常年的忠实拥护者，许多孕妇产品专卖店也在浑然不觉的情况下破产。浑然不觉的背景下，大数据正在推动一股强劲的商业革命暗涌，零售商们早晚要面对的一个问题就是：究竟是在浑然不觉中崛起，还是在浑然不觉中灭亡。

在消费者的需求呈个性化发展的大趋势下，笔者建议零售商应该学会收集、储存和分析大量的数据，并发挥出这些数据的价值。基于大数据的业务模型将主导零售业后十年的格局，大数据对打破零售业常规局面具有重要作用，其能够帮助零售商们筛选信息，迎接挑战，并且利用技术为客户提供解决方案。

10.2.5 【案例】上品折扣：用大数据做全渠道营销

春节、元宵节、情人节，随着各种节日的相继来临，网购礼品热潮让网络商家体验着一波又一波的狂欢盛宴。而在国内，最早用折扣吸引大众眼球的，不是淘宝和京东，而是一家线下的品牌折扣连锁店——上品折扣。

上品折扣（Shopin）是中国都市型百货折扣连锁店旗舰品牌，囊括 8 家实体店和一家电子商务网站上品折扣网。上品折扣主要以联营模式为主，并在逐步开展采购买手业务，目前合作的供应商达 3000 多家，几千个品牌在上品折扣的门店和线上进行销售，逐步形成了线上线下一体化的经营模式。

从电商热潮到全渠道营销这个过程中，上品折扣管理层也意识到所谓的大数据时代，IT 技术将扮演越来越重要的角色，不仅仅是电商业务，未来的数据管理、实体店、营销、会员体系全部都需要一个更智能的数据库去做支撑。

上品折扣于 2009 年率先在传统零售商中开始了 B2C 网上商城业务，在行业中逐渐处于领先的低位。2009 年 5 月，上品折扣旗下官方购物网站“上品折扣网”上线，如图 10-7 所示。2010 年 4 月，“上品折扣网”实现了由单店购物系统到多店购物系统的升级。



图 10-7 上品折扣网主页

经过三年多的时间，上品折扣就积累了大量商品数据。为了利用好这些数据为企业和供应商服务，使其不再沉睡，让数据说话，上品折扣开始了 BI 分析的探索之路，将其定义为 BDA（商业数据分析）。

上品折扣在构建 BI 系统时，遵循一个原则：节约投资，选择适合企业现阶段的产品，深入了解业务，规划业务模型，再到数据仓库的实现。BDA 系统主要的特色是，对在上品销售的 3000 多家供应商，近 3000 家品牌商品做了全品类、单品级的数据分析，对供

应商提供了实时的销售信息，这可以推动供应商更有效地补货，为商品规划部门、营销部门，以及电子商务提供了统一的数据分析。

目前上品折扣的 8 个卖场内有超过 1000 个知名品牌，每年卖家需要管理的 SKU（Stock Keeping Unit，库存量单位）数超过 300 万。上品折扣对卖场内所有商品都做了数据化管理，并通过替导购员配备 iPad 实现了线上线下实时信息传输和库存共享。例如，上品折扣会在品类、季节、适合人群、款式等角度对数据进行详细的分析。

专家提醒

受餐馆用 iPad 点菜的启发，上品折扣还开发了自主品牌 PDA 用于销售环节。上品折扣 PDA 主要用于解决商品数据的采集和现场的物品销售。上品折扣的 6000 名销售员，每个人的数据都是对接到同一个系统中。

上品折扣的 BDA 系统应用也是刚刚起步，从无到有需要一个适应和认知认同的过程，它正在起到积极的作用。上品折扣从 2012 年开始投入数千万元，与 SAP 合作改造数据系统，SAP 甚至把上品折扣列为亚洲的零售灯塔客户。这是一个庞大业务梳理的过程，涉及高管人才、数据库、供应链、品牌、终端等方方面面，整个改造花费了一年半时间。

这次 SAP 帮助上品折扣做出了两个突破，第一是同步线上线下的库存管理，这是一个根本前提。第二是在管理架构上借鉴欧美买手制度，以前联营做的多买手做的少，有很多不到位的地方，以后基于买手的自我搭建能力都将有很大变化。

随着消费者购物习惯改变，百货卖场也需要围绕顾客衍生出新的销售渠道。此外，上品折扣目前还通过微信进行营销。据悉，未来上品折扣还将借助二维码通过邮报和印刷品直接零售商品。

上品折扣希望通过实体门店持续拓展全渠道业务，在基于一整套的 IT 系统中，通过电商、移动互联网、BI、DM、电视购物甚至 Call Center（呼叫中心）等多个渠道发展。上品折扣对全渠道营销寄予厚望，希望未来 3～5 年内能够占到公司营收 50%以上。

【案例解析】：从本案例可以看出，从电商热潮到全渠道营销这个过程中，上品折扣管理层也意识到所谓的大数据时代，IT 技术将扮演越来越重要的角色，不仅仅是电商业务，未来的数据管理、实体店、营销、会员体系全部都需要一个更智能的数据库去做支撑。

另外，上品折扣借助移动互联网终端来加强用户体验，采用信息化技术和产品不仅可以不断地提升企业竞争力，还能满足消费者日益增长的购物需求。

10.2.6 【案例】阿迪达斯：用大数据带来利润

2009 年 8 月初，在国内大型的招聘网站上相继出现了一则阿迪达斯公司的招聘广

告，职位为 Inventory Sales Specialist（存货销售专员），工作地点为上海阿迪达斯中国区总部。该职位描述的首要条件是：能够按照不同渠道，根据实际库存情况，制定一个年度库存削减计划。

这样的招聘信息在平时并不会引起人们的关注，而在众多渠道商纷纷大面积低价清理手头存货，有人退出、有人倒闭，甚至有经销商干脆不去阿迪达斯处提货的情况下，它的意义就显得很不寻常。越来越多的事实表明，存货问题已经让阿迪达斯进入到一个危机之中，程度甚至让其难以控制，并将对其今年后两季甚至明年的发展造成影响。

阿迪达斯本应更早启用类似的专业库存管理人才来准确预期产能变化。但不久后，与阿迪达斯一起乐观地预期市场增长的渠道商们发现，由于市场并未达到预期，经销商多拿的货变成了自己身上的“沉重包袱”。

迫于经营压力，甚至有一些经销商因缺少资金宁愿违反协议拒不提货。由于阿迪达斯与经销商采取的是半年预订、货到付款的方式，这些未根据协议提走的、积压在阿迪达斯的仓库中的货品总款甚至高达上亿元人民币。

库存危机后，阿迪达斯从“批发型”公司转为“零售驱动型”公司，它从过去只关注把产品卖给经销商，变成了将产品卖到终端消费者手中的有力推动者。而数据收集分析，恰恰能让其更好地帮助经销商提高售罄率。

阿迪达斯产品线丰富，过去，面对展厅里各式各样的产品，经销商很容易按个人偏好下订单。现在，阿迪达斯会用数据说话，帮助经销商选择最适合的产品。

（1）抓牢不同区域的消费者需求：一、二线城市的消费者对品牌和时尚更为敏感，可以重点投放采用前沿科技的产品、运动经典系列的服装以及设计师合作产品系列；在低线城市，消费者更关注产品的价值与功能，诸如纯棉制品这样高性价比的产品，在这些市场会更受欢迎。

（2）分析不同区域的经销商数据：阿迪达斯会参照经销商的终端数据，给予更具体的产品订购建议。例如，阿迪达斯可能会告诉某低线市场的经销商，在其辖区，普通跑步鞋比添加了减震设备的跑鞋更好卖；至于颜色，比起红色，当地消费者更偏爱蓝色。

推动这种订货方式，阿迪达斯得到了经销商们的认可。一方面降低了他们的库存，另一方面增加了单店销售率。卖的更多，销售率更高，也意味着更高的利润。挖掘大数据，让阿迪达斯有了许多有趣的发现。例如，同为一线城市，北京和上海消费趋势不同，气候是主要的原因。实际上，对大数据的运用，也顺应了阿迪达斯大中华区战略转型的需要，如图 10-8 所示。

下面看一位阿迪达斯的忠实经销商是如何利用大数据渡过危机并走向成功的。

2012 年 12 月，厦门育泰在福建省泉州市的一个沿海县级市——南安开出了一家新店。南安算上周边地区也有 150 万人口，它一般会被定义为中国这个庞大市场里的四线或五线城市。

厦门育泰是阿迪达斯在福建最大的经销商，当厦门育泰把第一家阿迪达斯门店开到

南安的时候，南安还只有一个购物中心，门店第一年的利润是 12 万元。现在，随着另一个受到年轻人欢迎的购物中心的建起，育泰公司也挑选了一个临街的好位置，开出了这一家 120 平方米的新店。



图 10-8　阿迪达斯的大数据战略目标

厦门育泰总经理叶向阳看着同行大多仍身陷库存泥潭，他庆幸自己选对了合作伙伴。他的厦门育泰贸易有限公司与阿迪达斯合作已有 13 年，如今旗下已拥有 100 多家阿迪达斯门店。他说，“2008 年之后，库存问题确实很严重，但我们合作解决问题，生意再次回到了正轨。”

现在，叶向阳每天都会收集门店的销售数据，并将它们上传至阿迪达斯。收到数据后，阿迪达斯对数据做整合、分析，再用于指导经销商卖货。研究这些数据，让阿迪达斯和经销商们可以更准确了解当地消费者对商品颜色、款式、功能的偏好，同时知道什么价位的产品更容易被接受。

叶向阳的生意也在过去两年中有了巨大变化，在他目前经营的总共 100 多家门店中，有 50%都位于像南安这样的四五线城市。

【案例解析】：在本案例中，阿迪达斯通过与经销商伙伴展开了更加紧密的合作，以统计到更为确切可靠的终端消费数据，有效帮助自己重新定义了产品供给组合，从而可以在适当的时机，将符合消费者口味的产品投放到相应的区域市场。简而言之，阿迪达斯还只是利用数据对客户进行细分，然后开展针对性营销。

不过，笔者认为阿迪达斯还缺乏创新能力，想要创新就必须要学会利用大数据的“预知”能力。零售企业可以利用大数据事先捕捉顾客的关注点和需求，并且给出可执行的解决方案，帮助回流客户。另一方面，社交媒体、电子商务、物联网等新应用的兴起，打破了企业原有的价值链围墙，仅对原有价值链各个环节的数据进行分析，已经不能满足需求，零售企业需要借助大数据战略打破数据边界，了解更为全面的运营及运营环境的全景图。

读书笔记



11

制造：更快更好地生产

学前提示

围绕大数据的话题主要集中在点击流数据、倾向性分析和消费者定位。但其实在大数据背后，机器到机器的通信以及先进的分析功能可能会完全改变我们周围的世界。本章将介绍大数据在传统生产制造业的解决方案和应用案例。

要点展示

- 生产制造业大数据解决方案
- 生产制造业大数据应用案例

11.1 生产制造业大数据解决方案

如今，大数据正处于引爆点，有数十亿美元投入到将海量信息转化为对商业有价值的洞察力。不过，大数据的内涵不仅仅在于数字和洞察力，它对促进智能化生产也有着重大意义。

11.1.1 大数据对生产制造业的影响

笔者认为，对于大数据的理解不仅是其中存在的价值，而更在于可以进行种种连接以赋予大数据主动性和预测性——或让信息智能化。

然而，为了使信息智能化，新的连接需要建立起来，这样大数据才能“知道”何时以何种方式前往何地。大数据看起来可能像是工作流程的一种简单升级，但事实上，它代表的东西可能是自工业革命以来意义最深远的商业和技术的融合——工业互联网（Industrial Internet），如图 11-1 所示。



图 11-1 工业互联网（Industrial Internet）

工业互联网将整合两大革命性转变的优势：其一是工业革命，伴随着工业革命，出现了无数台机器、设备、机组和工作站；其二则是更为强大的网络革命，在其影响之下，计算、信息与通信系统应运而生并不断发展。

工业互联网是指全球工业系统与高级计算、分析、感应技术以及互联网连接融合的结果。它通过智能机器间的连接并最终将人机连接，结合软件和大数据分析，重构全球工业，激发生产力，让世界更美好、更快速、更安全、更清洁且更经济。伴随着这样的

发展，工业互联网的 3 种元素逐渐融合，充分体现出它的精髓，如图 11-2 所示。

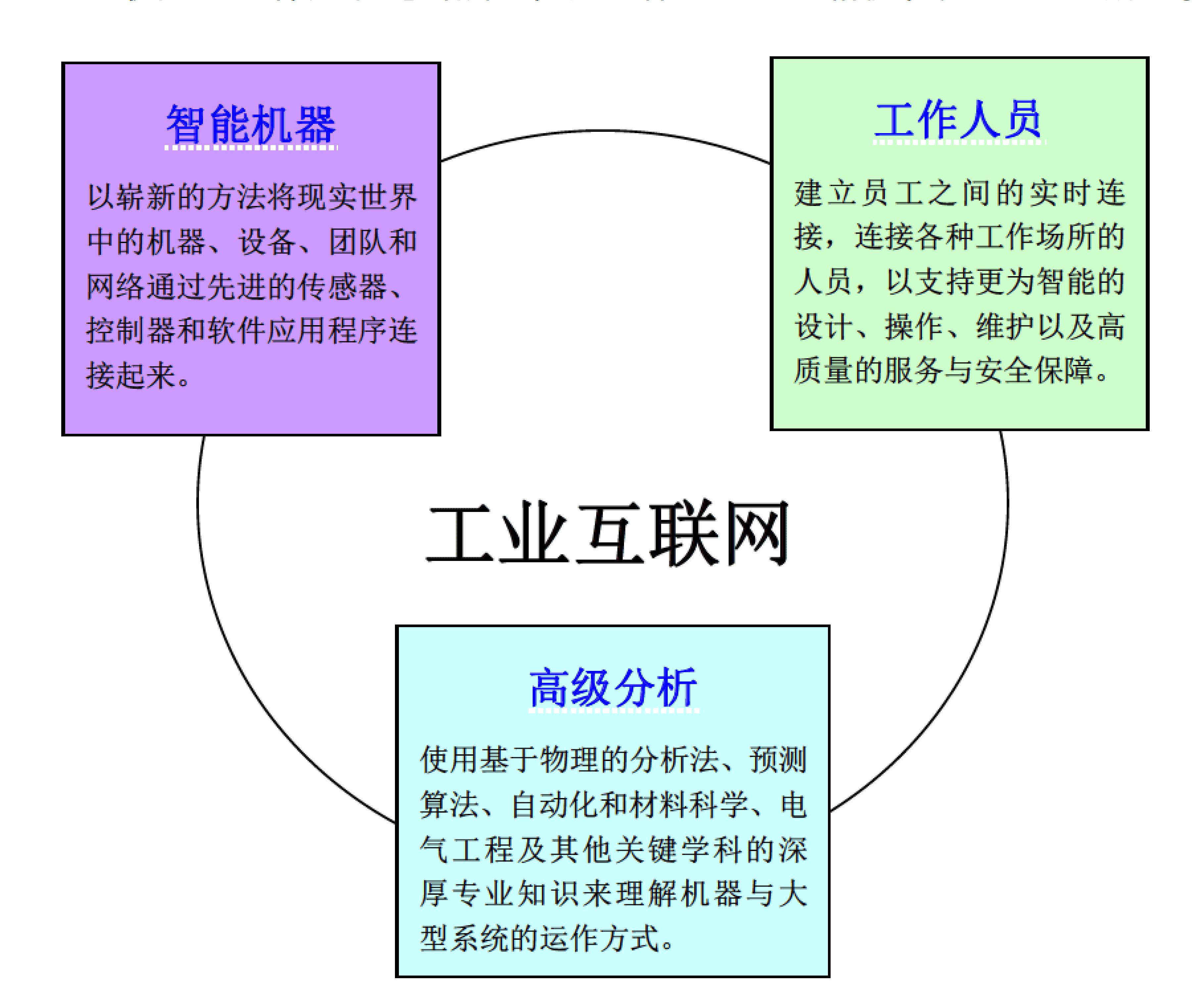


图 11-2　工业互联网的 3 种元素

工业互联网将这些元素融合起来，将为企业与经济体提供新的机遇。例如，传统的统计方法采用历史数据收集技术，这种方式通常将数据、分析和决策分隔开来。伴随着先进的系统监控和信息技术成本的下降，工作能力大大提高，实时数据处理的规模得以大大提升，高频率的实时数据为系统操作提供全新视野。

大数据是工业互联网的命脉，但工业互联网同样意味着开发新的软件和分析方法，以便从原先不存在连接的地方——如机器内部——提取和厘清数据。通过让机器经由软件连接到互联网，数据由此产生，数据洞察不断积累，但更重要的是，这些机器现在成为一个紧密结合的智能网络的组成部分，这个网络被构建用来让关键信息实现安全的自动化传输，以对性能问题进行预测。这意味着及时省下来的数千亿美元和各大行业可利用的资源。

例如，很多时候停电事故得不到修复，有时长达数周时间，这是因为线路断开的地点无法被立刻获知，或是因为系统需要进行大规模的检修而发生故障的部位可能位于世界的另一侧。然而，在工业互联网中，从发电的巨大机器到电线杆上的变压器，一切都可以连接到互联网上，从而提供状态更新和性能数据。由此，维修人员可以在潜在问题造成公司损失数百万或数十亿美元，并在浪费客户时间之前抢先采取行动，他们将能够

预测哪儿出了错，并准备好修复所需的零部件。

工业互联网的应用能够帮助中国的航空、电力、铁路、医疗、石油天然气等主要行业实现生产率提升 1%，在未来 15 年将有潜力让这些行业节省成本约 240 亿美元。当然，要做到这些，不仅需要充分利用大数据，还需要建立正确的连接让大数据为我们服务。

专家提醒

机器分析为分析流程开辟新维度，各种物理方式的结合、行业特定领域的专业知识、信息流的自动化与预测能力相互结合可与现有的整套“大数据”工具联手合作。最终，工业互联网将涵盖传统方式与新的混合方式，通过先进的特定行业分析，充分地利用历史与实时数据。

11.1.2 生产制造业如何利用大数据

如今，大数据已经带来以下场景。

场景 1：通信公司可以根据你习惯阅读手机报的时间来不断调整发送时间。

场景 2：午餐时，餐厅也会分析你的偏好和需要来管理和优化原料的供应。

场景 3：超市也会根据商品销售的关联分析来不断调整货架，让你更容易发现和购买所需的商品。

这些都是大数据时代的典型商业智能应用。机械制造业是最早开始走上信息化道路的行业之一，其业务信息化系统已经趋于完善，而随着业务系统的完善，也随之带来了一个问题，以 TB 级增长的数据如何“消化”，如何让这些数据返过来促进业务的创新？

笔者认为，在大数据时代，全球工业系统与高级计算、分析、传感技术及互联网将会进行高度融合。工业互联网将利用数据来连接智能机器，并最终将人机连接，结合软件和大数据分析，重构全球工业，激发生产率，让世界发展更快速、更安全、更清洁且更经济。

实际上，很少有企业是因为单纯的积累数据而了解大数据，更多的动力依然是来自业务需求，也就是利益的需求。大数据分析可以让机械制造业的各个部门的数据得到充分的利用，如表 11-1 所示。

表 11-1 大数据分析在机械制造业各个部门的应用

企业部门	主要应用
财务部门	财务部门可以牵头建立成本控制体系；生产部门可以牵头建立 KPI（Key Performance Indicator，企业关键绩效指标）体系；以及信息管理部门牵头建立数据仓库，支持 KPI 体系和成本控制体系等的平台；还有人力资源、供应链等各个部门都可以在已有的数据上做出更多的业务创新

续表

企业部门	主要应用
生产部门	生产部所要解决的问题不仅是对流程、业务、订单、事务等的规范化管理，还要对产生的数据进行进一步的分析，以进一步实现业务流程优化。如今，很多企业都在强调创新、高效，但如果没有一个统一的数据分析平台，生产部门就依然会陷入处理各种报表的琐碎业务中，没有时间去考虑创新和高效。因此，利用数据分析平台不仅能够连接各类主流数据库，还可以支持多种数据来源，保证了数据分析的完整性，再利用多种数据分析手段挖掘数据的价值，从而让生产部门发挥出更大的创新价值
信息部门	信息部门需要一个支撑的平台，这类需求是明显的商业智能的需求，需要利用大数据分析产品来实现对于多业务系统数据的整合，同时根据各业务部门的需要定制报表，通过条件参数来实现自动刷新报表数据的功能。大数据分析平台能够与各业务平台进行良好的集成应用，这样可以为企业量身定制辅助决策体系，以图表并举的方式将全面的数据分析结果呈现给管理者，也可以免除基层工作人员大量的手工工作，同时也能及时、准确地将数据以各部门所要的形式呈现出来

事实上，无论是哪个领域的应用，都是通过对多维数据库的旋转、切片、钻取、多维度切换等手段进行分析，从而使各管理人员或业务人员能够真正将主要精力从“手工劳动”生成报表或报告转移到应用先进的手段去发现问题，解决问题上来。

专家提醒

信息化产业的关键是从许多来自企业不同的运作系统的数据中提取出有用的数据并进行清理，以保证数据的正确性，然后经过抽取（Extraction）、转换（Transformation）和装载（Load），即 ETL 过程，合并到一个企业级的数据仓库里，从而得到企业数据的一个全局视图，在此基础上利用合适的查询和分析工具、数据挖掘工具、OLAP 工具等对其进行分析和处理（这时信息变为辅助决策的知识），最后将知识呈现给管理者，为管理者的决策过程提供支持。其中，ETL 是负责完成数据从数据源向目标数据仓库转化的过程，是实施数据仓库的重要步骤。

11.2 生产制造业大数据应用案例

中国是制造大国，但还不是制造强国。目前，我国制造业的持续发展面临诸多问题。例如，资源环境的制约异常突出，产业发展乏力，产业技术创新能力薄弱，产业结构调整的任务非常艰巨，发展方式转变十分困难。要实现由制造大国向制造强国的转变，加快发展先进制造业势在必行，而大数据就是最好的帮手。本节主要介绍生产制造业大数

据的应用案例，希望对读者有一定的启发和学习价值。

11.2.1 【案例】大数据结合 ERP 助力生产

笔者的好友王贵是一家家具生产公司的老板，在笔者刚接触到“大数据”这一概念时，就曾与他公开交流过大数据的应用方式。如今，王贵在操心业务的同时，也知道必须要更好地利用信息化系统，这样才能更好地完成任务。

近日，王贵正在想办法提高各生产线的效率，使计划生产达到 80%而不是现在的 60%。要达到这一目的，王贵首先必须知道各个生产线的生产状况，然后可以随时对生产线做出调整。其实王贵也没有想让所有生产线满负荷运转，因为他很清楚那是无法实现的。

但计划生产达到 80%，甚至再低一点 70%是完全可以实现的，而且也会在很大程度上提高生产效率，从而为企业增加利润。

据笔者了解，王贵所在的企业是一家典型的多品种、小批量、根据订单生产的生产制造型企业。两年前，颇具科技头脑的王贵就开始应用企业资源计划系统（Enterprise Resource Planning，ERP）建立起了合理、高效的生产计划编制体系，消灭了信息孤岛，基本实现了数据共享。另外，王贵利用该系统使生产、采购、销售、库存等环节连接成了一个整体，这在很大程度上解决了以往由于信息不匹配造成的影响，甚至是经济损失。

虽然企业利用 ERP 系统解决了很多采购、生产等环节出现的问题，提高了订单交付的及时率、准确率，同时也提高了客户的满意度。但是，王贵还是忧心忡忡，主要是因为企业的品种多而杂，而且订单随时性很强，经常会出现临时加单、订单调整的情况，让企业措手不及。同时，王贵还要清楚地了解退货情况，具体原因是什么等信息，这让他的工作量不断加大。

另外，王贵还要想办法掌握一些重要信息，例如，该通过什么样的方式了解订单的趋势，提前做好准备；通过多种维度去分析退货的情况和原因，同时采取措施降低退货率。平时，这些信息也要花费很大的精力和很长的时间去统计，让本来就繁重的工作又增加了更繁琐的工作内容。

【案例解析】：笔者认为，如果企业规模还很小，几个人当面沟通就能搞清楚全部状况时，可能不会需要 ERP 系统。但除非企业不想再继续成长，否则从整体策略的角度，重新规划企业资源运用方式与营运模式，并据此导入 ERP 系统，并顺势采取合理化、标准化的步骤，会是任何一个有追求发展的企业管理者的必然选择。

ERP 系统是事务性处理系统，它解决了多个子系统之间的数据流转的问题，每个环节的工作人员通过处理不同的单据来记录整个过程所发生的数据。然而，ERP 系统却也存在一些不足之处：

➢ 数据深层次的信息却没有被挖掘出来。

- ERP 系统的操作大多面向基层人员，以业务操作为主。
- 如果作为管理者使用，ERP 系统的易用性又不够，而他们又是最需要利用数据来进行辅助决策的。

在本案例中，对于像王贵这种中层甚至高层使用者来讲，需要为他们提供一套操作简单、内容全面的数据分析平台，作为 ERP 系统和管理者之间的桥梁，如图 11-3 所示。只有利用这样的系统，才能让他们摆脱目前的状况，最好的办法就是通过信息化的方式来帮助他们去完成这些工作，提高工作效率，也为决策提供依据，从而也使他们有更多的时间和精力去研究部门现状，剖析企业问题，从而更好地实现创新发展企业的目的。

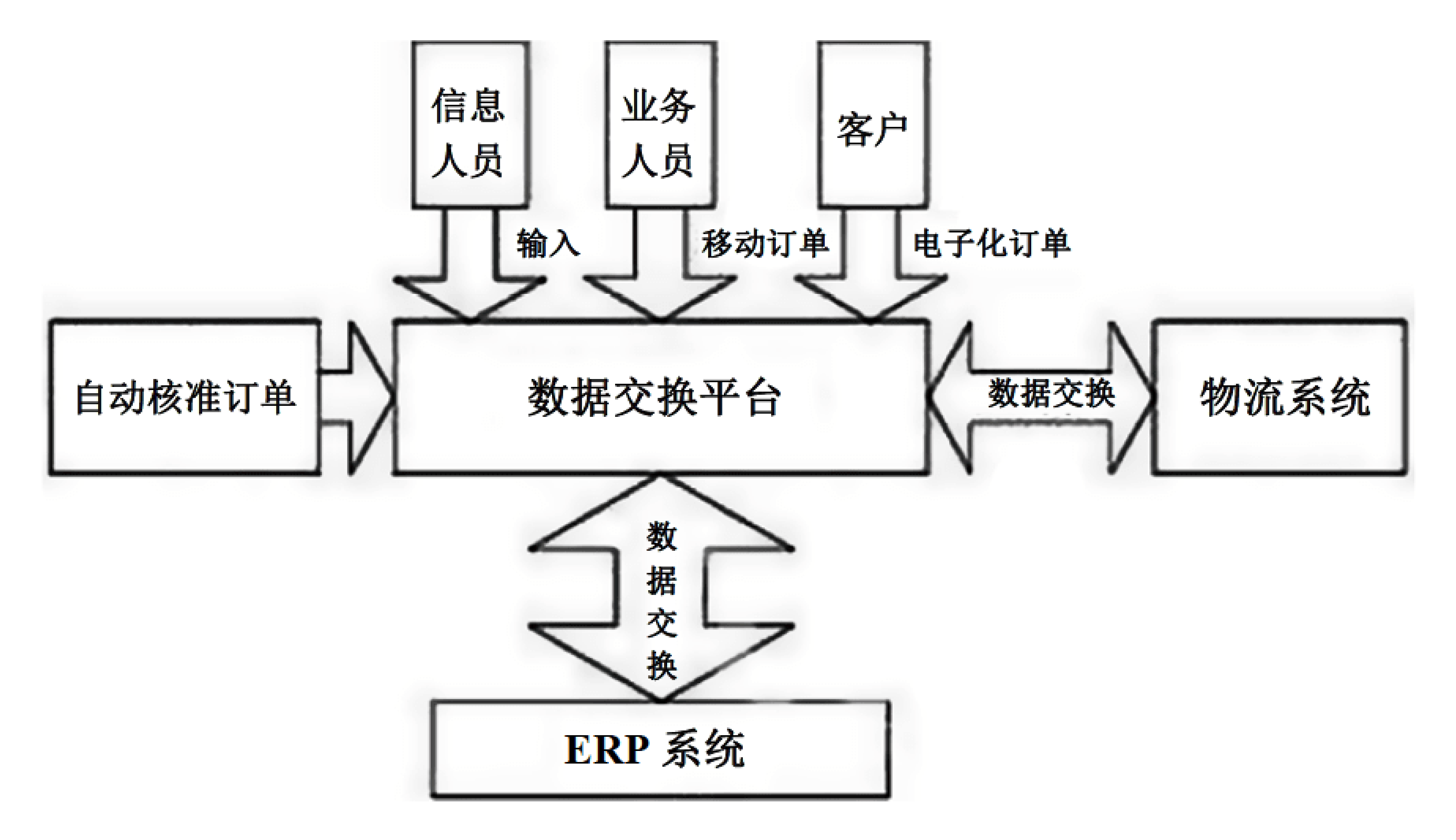


图 11-3　用数据连接各个系统

同时，笔者预计，在未来的几年内，将会有为数不少的企业进入大数据市场，这个市场的竞争也将更加激烈。

11.2.2　【案例】大数据改变福特汽车的制造

过去十年，福特公司经历了一个非常困难的时期，当时福特失去了近半数的员工，整个公司濒临灭亡。因此，早在 20 世纪 90 年代，福特公司就已经开始认真考虑是否使用数据分析工具，当时服务器和存储越来越便宜，很多华尔街的公司都在向世界展示利用数据建模可以实现什么。

福特公司内部开始出现各种分析小组，包括 Ginder 研究中心小组、市场部单独的小组、福特信贷（FordCredit）部门的小组。尽管如此，所有这些分析小组都把精力集中在一些非常具体的任务上，例如福特信贷部门的风险分析，或者像研究中心那样做更为抽象的科学工作，而且这些都被称之为核心的业务驱动力。

福特公司的大数据分析负责人 John Ginder 在福特研究中心（Ford Research）管

理着系统分析和环境科学（Systems Analytics and Environmental Sciences）团队。与此同时，另一个因素开始发挥作用——新任 CEO 的到来。2006 年，新的首席执行官 AlanMulally 来到福特，每周他都要与手里拿着各种图表的直接汇报人开会，并经过层层细化，鼓励公司内部使用数据驱动的方法，他影响了整个福特的管理文化。

福特的另外一个重要的大数据资产来自福特产品开发流程和产品本身产生的大量有用数据。福特内部产生的大量数据，包括来自业务运营、汽车产品研究活动以及互联网上的客户数据，所有这些数据对于福特来说意味着巨大的商机，但是福特需要新的专业技术和平台来管理这些数据。福特的研究部门正在测试 Hadoop 系统，试图整合手头拥有的所有数据源。

福特的制造工厂以及汽车产品都安装了各种测量仪表，它们都是闭合的控制系统。每辆汽车中也安装有大量传感器，但目前这些数据都还停留在汽车内部，但是 Ginder 认为采集这些数据，包括车辆运行状况和消费者操控汽车方式的数据，并将这些数据分析后反馈给设计流程将非常有助于优化用户体验。

除了采集结构化数据进行分析外，福特还将触角伸向了非结构化的消费者情报数据。虽然不少财富 500 强企业也在进行类似的社会化分析，但是福特分析 Web 非结构化数据的方法与众不同，该方法甚至能够影响到公司对汽车产品销售业绩的预测。例如，福特使用 Google Trends（如图 11-4 所示）来监测搜索关键词的流行度，帮助企业做出内部销售预测。

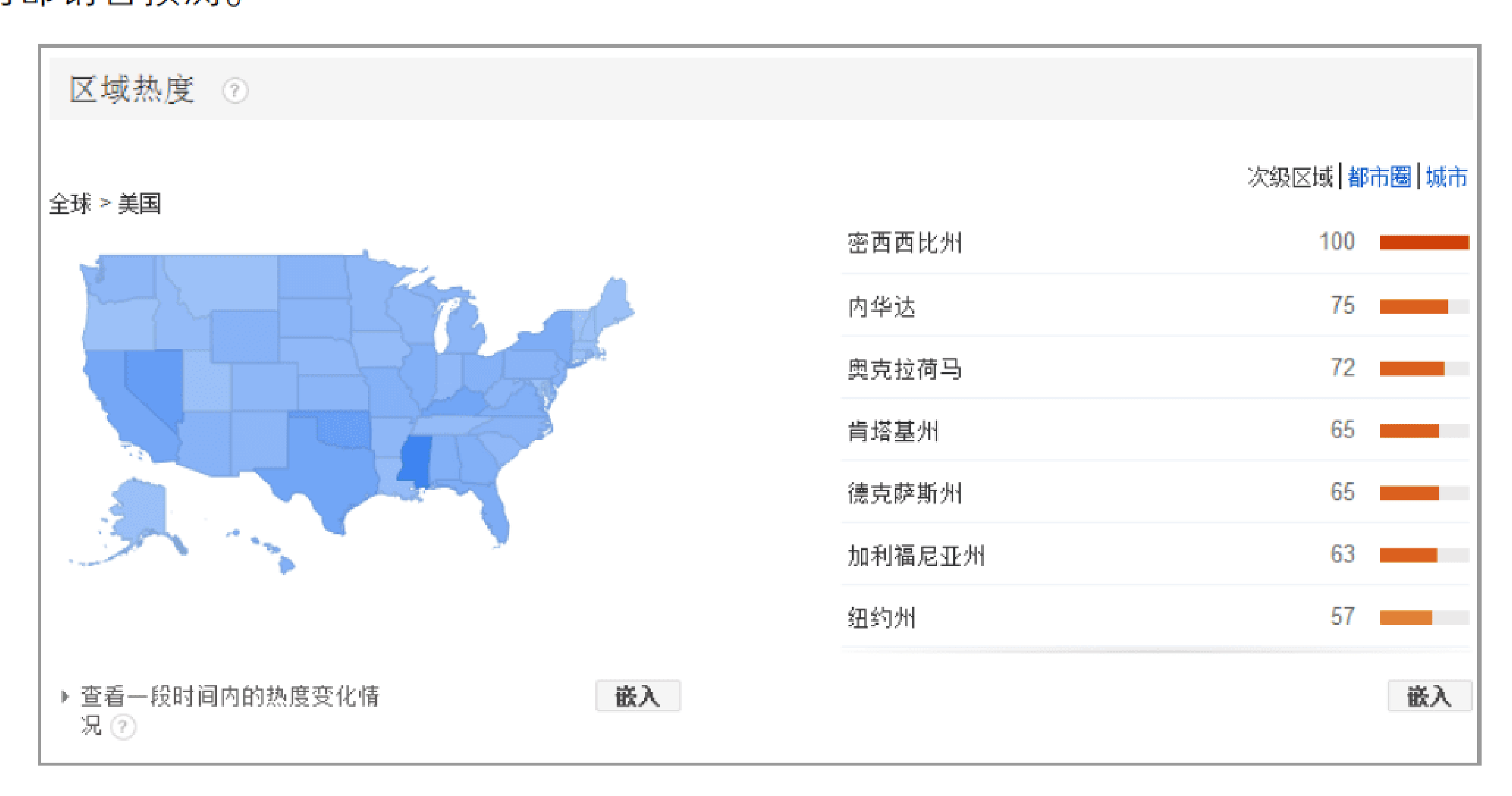


图 11-4 Google Trends 的区域热度分析

在 Ginder 的眼里，福特的大数据分析还只是“皮毛功夫”，因为大数据分析工具目前并不成熟。虽然能够洞见大数据的未来，但是 Ginder 认为现实和未来还有相当的落差。“大数据的未来很美妙，不过我们现在的问题是专业人才和工具都很缺乏。虽然我

们有自己的专家，可以利用目前的大数据工具开发一些大数据应用解决具体业务问题。但是将来我希望能把大数据分析扩展到所有数据，届时数据专家——而不是电脑专家，能充分发掘大数据的商业价值。”

在 Ginder 的眼里，福特的大数据未来还意味着数据的开放，福特将与开源社区大量分享自己的数据，造福社会。不久前福特的硅谷实验室（SVL）正式揭幕，其定位是“大数据、开源创新和用户体验”。现在，分析已经深入福特公司的文化当中，大数据分析的兴起，为这家汽车制造商带来了全新的机遇。

专家提醒

例如，福特产品开发团队曾经对 SUV 是否应该采取掀背式（即手动打开车后行李箱车门）或电动式进行分析。如果选择后者，门会自动打开，便捷又智能；但这种方式会出现车门开启有限的问题。此前采用定期调查的方式并没有发现这个问题，但后来根据对社交媒体的关注和分析，发现很多人都在谈论这些问题。

【案例解析】：当问起汽车的制造过程，大多数人脑子里随即浮现的是各种生产装配流水线和制造机器。然而在本案例中，福特在产品的研发设计阶段，大数据就已经对汽车的部件和功能产生了重要影响。

笔者发现，福特目前主要依赖开源工具如 Hadoop 来管理大数据集，并通过 R-Project（另外一个开源数据分析工具）来进行统计分析，此外数据挖掘和文本挖掘使用的也都是开源工具。

虽然开源大数据工具非常强大，可扩展性也很好，但是只有高水平的数据分析专家和程序员才能使用。此外，大数据的一个大趋势是，非技术人员也将能通过自然语言工具访问大数据集。未来的“数据科学家”不是懂得如何书写合乎规范的 SQL 查询语句的人，而是知道如何提出正确问题的业务分析师，只有他们能够发现影响公司决策的“数据珠宝”。

11.2.3 【案例】长安汽车数据与制造的结合

长安汽车充分把握了我国西部大开发和 WTO 带来的双重发展机遇，以先进的信息技术全面提升了公司的信息技术应用水平，锻造出企业的核心竞争力。作为中国汽车行业的排头兵之一，长安汽车的信息化建设同样有不俗的表现。从 20 世纪 90 年代初踏上信息化之路以来，如今长安汽车已经在研发、生产、销售各个环节应用了信息化系统，实现了信息化对业务的全面支撑。

2000 年是长安汽车信息化建设的一个分水岭。随着国内车市空前繁荣，长安的面前既有危及存亡的竞争压力，又有跳跃式发展的巨大商机。在这种机遇与挑战下，汽车企业必须采用全球化的、灵活的电子商务供应链模式，通过完整、集成的信息化平台强化

汽车企业在速度、创新、出色的客户关怀以及整个供应链协同方面的突出能力，从而在本土竞争中立于不败之地，并谋求在世界范围内做大做强。同时，因为业务战略的转变，长安汽车此前建立的 38 个不同的信息系统开始无法满足新的需求。为了长安汽车未来的发展，企业高层决定将信息化迁移到更大的平台。

2001 年，长安汽车与 Oracle（甲骨文）公司确定了战略合作伙伴关系，应用了 Oracle 的 ERP、e-HR、CRM 等系统，并与 Oracle 的支持服务部门建立了长期的合作伙伴关系。通过与 Oracle 的合作，让长安汽车更加确信采用“一线贯通”的方式建设信息化符合其战略发展方向。

长安汽车在国内外有众多的产业基地、分/子公司，对应的信息系统相当庞大。在现代企业竞争中数据的力量不容小觑，信息系统里流淌的数据，对于企业来说如同人的血液一样重要。总结起来，应该说长安采用的是“一线贯通”的方法来实现企业信息化，其对未来的企业平台架构、运营成本及推进一体化管理都有很好的用处。

长安汽车董事及副总裁马军使用最多的一个词也是“数据”。他认为，“作为一个企业，重要的是你知不知道你下面的数据，知不知道数据形成的业绩与竞争对手的数据差异在哪里。”

在以往，长安所有产品的开发数据、工业数据、制造数据由不同部门各自分管，导致从研发到生产数据并不唯一，系统之间的关联性也不强。为此，长安建起了一套以 PDM 系统为核心的全球在线研发平台，把数据源打通，使所有数据在同一个链条上互动，优化了在线协同研发机制。

长安汽车使用信息系统之后，企业得到的是一个数据链，从原来点的数据到线的数据到一个数据链的数据。有了数据链，企业可以系统地去和竞争对手，和行业进行比较，甚至和国外的先进的汽车企业作比较。如果没有这些数据，企业在做竞争策略时只能凭感觉和经验。

2010 年，长安汽车预计产销汽车 185 万辆以上，销售收入达到 1 千亿以上，这其中信息化功不可没。对于成本控制、物料管理、差异分析和风险分析，信息化发挥了重要的角色。长安汽车正在按照新的发展规划，部署新的 IT 运营，来配合业务的快速成长和发展需求。

信息化本身是一个持续不断的过程。在这个过程中，不断有问题出现并需要解决。在取得了诸多成绩的同时，长安汽车信息化同样面临着挑战。

目前，长安汽车挑战来自以下两个方面：

（1）如何保证信息的安全与共享？随着软件应用越来越多，运用范围越来越广，信息安全成为一个重要问题。长安汽车信息系统采用的是集中管理的方式，如果发生信息泄露就是大问题。而如果不集中管理就不能共享，不能共享则造成成本升高以及绩效评价的不公平。这一矛盾对长安汽车信息化形成了挑战。

（2）如何保障 24 小时的运营？对于这个问题，长安汽车目前是“两地三灾备”的策略，即同城有两个灾备中心，异地有一个灾备中心。如何在全国 11 个城市去部署运用，是长安汽车的另一大挑战。

通过数据平台，长安汽车解决了全球共享单一数据源、提供实时准确的数据、支撑五国九地、7×24 小时在线协同研发等问题。同时，通过数字化设计和制造仿真分析，提前发现问题，以减少后期变更成本，减少实物验证次数。在抓住了数据源之后，长安信息管理部把研发部分的成本控制在了原来的 80%上下，协同效率的提升更使得生产等环节的成本得到控制。

专家提醒

很多企业的数据是不对称的，如果数据在流转的过程中出现人为加工修改，就会为企业决策带来很大的潜在风险。拥有数据，才能与竞争对手对比。信息系统保证了数据的透明与规范，让数据呈现在所有应该共享的人面前。

首先，信息系统带来了数据的对称，同一系统中授权一致的人会看到相同的信息，谁也没办法隐藏信息，它是透明的。同时，数据的对称规范了管理，如果没有数据，想做到精益管理基本是空谈。当然，做到数据的透明规范与共享，最终的目的还是实现企业整体效率的提升。

2010 年 10 月 31 日，长安汽车发布了全新的品牌标识，并宣布 2020 年的战略目标是实现年产销 600 万辆，成为世界级的汽车企业。在 2013 年上半年，长安集团实现了营业收入 197.51 亿元，同比增长达到 40.63%，其中汽车制造业务整体的毛利率达到 16.43%，比 2012 年同期提高了 0.9%，而产品毛利的提升主要就来自于产品结构的优化及持续的成本控制。

【案例解析】：信息化建设对于现代化企业来说是一场挑战，而这场挑战的核心内容便是数据应用。越来越多的企业开始重视以数据为核心的信息化建设整合，其中数据恢复被认为是最重要也是最容易被忽视的环节之一。

在本案例中，总的来看长安汽车选择 Oracle 是因为 Oracle 在互联网领域的成功经验和数据库基础、引领行业的技术把握能力和前瞻性以及良好的品牌形象与服务。Oracle 是第一家将应用软件产品向互联网演进的软件公司，全球财富 500 强中，96%的企业都不约而同地采用了 Oracle 大数据解决方案，以 Oracle 技术产品和解决方案作为信息系统建设的标准。

另外，长安在建立电子商务交易平台和营销方面，也是借助了 Oracle 领先的技术优势和丰富的实践经验。总之，数据信息为“虚”，生产制造为“实”，“虚”、“实”结合推动着长安集团的管理提升和成本控制。笔者也拭目以待，看长安汽车与 Oracle 继续长远而密切的合作，将创造更多的收益，获取更长远的发展。

11.2.4 【案例】乐百氏 BI 系统助力企业成长

乐百氏集团是闻名全国的大型食品饮料企业，中国饮料工业十强企业之一，公司目前在全国的布局为 5 个事业部，数千销售人员，管理全国约 300000 个销售终端。

2006 年，随着乐百氏的“战线”发展越来越长，业务员提交的销售报表格式越来越繁杂，需要投放促销资源的点越来越多，集团公司管理层的脑袋也随之越来越大。成立信息化部门以及构建 BI 系统迫在眉睫。

2006 年 2 月，乐百氏挑了几名 IT 助手，拉起一支全职的项目队伍，开始跑分公司进行需求分析，准备建立一套完善的市场分析系统。由于必须先从市场上拿到指定的数据，才能用于数据分析，乐百氏决定与明基逐鹿合作来完成企业 BI 系统的构建，并将项目分为数据采集与数据分析两个阶段。

专家提醒

明基逐鹿（BenQ Guru）是中国领先的 IT 技术、顾问服务、业务流程外包解决方案提供商，旨在将明基集团 20 多年全球管理运营经验与在数百家知名企业累计的管理真知，通过 greenOffice、eHR、SCM、MES 规划实施及 IT Service 分享给国内快速成长的企业客户。

1. 数据采集

在乐百氏的经营过程中，数据采集的关键指标很明确，如各销售网点的销售情况、库存情况、大超市的各项费用等，这些数据通过分公司或办事处录入到系统中，定时回传到总部，然后由明基逐鹿把这些数据制作成指定的分析报表。

逐鹿商业智能解决方案为其深度分销体系的监控与管理提供了重要保障。方案实施后，无论是渠道、组织、人员、终端、终端销售状况、市场状况、费用状况、库存状况、客户状况等信息，都能够通过企业绩效管理门户实时查询与分析，辅助管理者将“以售点为本”的渠道管理策略执行到位。

通过 BI 系统，乐百氏总部管理层可以轻易调出零售店的数据、经销商的数据，了解各分店的进货量、销售代表业绩及产品市场表现。

2. 数据分析

提高系统的任何一点适应性都需要借助人力，为此，乐百氏和明基逐鹿制定了一个共同目标：让系统更好用一些，让数据在所有区域经理面前显得更真实一些。

乐百氏 BI 系统数据分析的核心工作是设计报表系统与逻辑。

（1）利用报表系统，分析数据做出决策。报表系统是综合性的一套报表，需要将数据信息全方位展现出来，并通过报表系统将销售信息分成多种维度，穿插分析。这样，企业高层不仅可以看到该区域的总销量，还可以知道哪个客户销售比例更高，如果这个客户连续几个月都排在销售前三名，那么这个客户将是重点客户，可以让销售代表更多

地关注他。

（2）利用数据核对，提升 BI 系统适应性。BI 系统上线前的最后一步是核对数据，但这也是一个相对费力的过程。数据核对的难处，一方面在于 BI 所产生的报表是多步运算得到的结果而非简单的汇总，因此每个环节的计算都要反复核对；另一方面，数据核对不仅涉及项目组成员，企业各部门员工都得参与其中。这两个问题需要大量的人力去排查，从数据源的输入，到数据的传输、数据计算逻辑、数据展现，每个环节都不能忽略。经过大量的数据核对，BI 系统的适应性得到进一步提升。

3. 应用成果

目前，乐百氏的 BI 应用已经在各个分公司全面上线，BI 系统产生的效果逐渐显示出来。总的来说，BI 系统为乐百氏带来了以下 3 大好处：

（1）对于分公司的基层销售人员来说，利用 BI 系统，他们可以自己比对每人的销售业绩，总部对区域销售业绩的评判很少再引发异议；从中发现不符合要求的员工，可以毫不犹豫地把他开除，以此保持企业的快速增长。

（2）对于销售点的铺货来说，BI 系统统一了铺货率等数据上报的标准，规定铺货必须结合销售点的产品品种规格、陈列配合、季节性特征等因素，确定销售点可以接受的最大铺货量。无论是总部还是各区域经理，看到的数据是一致的、及时的，从而提高了办事效率。

（3）对于企业高层来说，BI 系统方便了企业高层激励考核基层组织，同时它也为企业的管理、决策及预测提供了数据依据。

【案例解析】：在本案例中，乐百氏手中掌握了海量的数据，同时拥有多种工具来移动、分析和发送这些数据，为企业的生产管理、人事管理以及营销策略带来极大的帮助。

在中国由世界制造中心转型成为世界上最大的消费市场的过程中，中国企业必须像乐百氏一样，挑战这样的一些转型包括供应链管理、产业链管理、上下游企业之间的关系以及如何应对仓储、物流的变化等。笔者认为，中国的人口、互联网用户数及移动互联网用户数都居全世界第一，在大数据时代，可供收集的数据将不再是瓶颈，关键之处是如何树立起应对大数据的意识，抓住这个机遇。

11.2.5 【案例】大数据可以破解“猪周期”

如果你每天去菜市场买菜，肯定会发现，近段时间以来，猪肉价格持续下跌。肉价跌了，市民的菜篮子是变轻了，但生猪养猪户的心情却变得沉重了，因为持续下跌的生猪收购价让部分养殖户利润受损。那么，我们该如何正确看待当前的肉价持续走低呢？生猪价格低位运行，养殖户又该如何规避风险呢？一会儿猪价太高了，CPI 上涨百姓生活受到影响，到底该如何才能解决“猪周期”难题（如图 11-5 所示）呢？

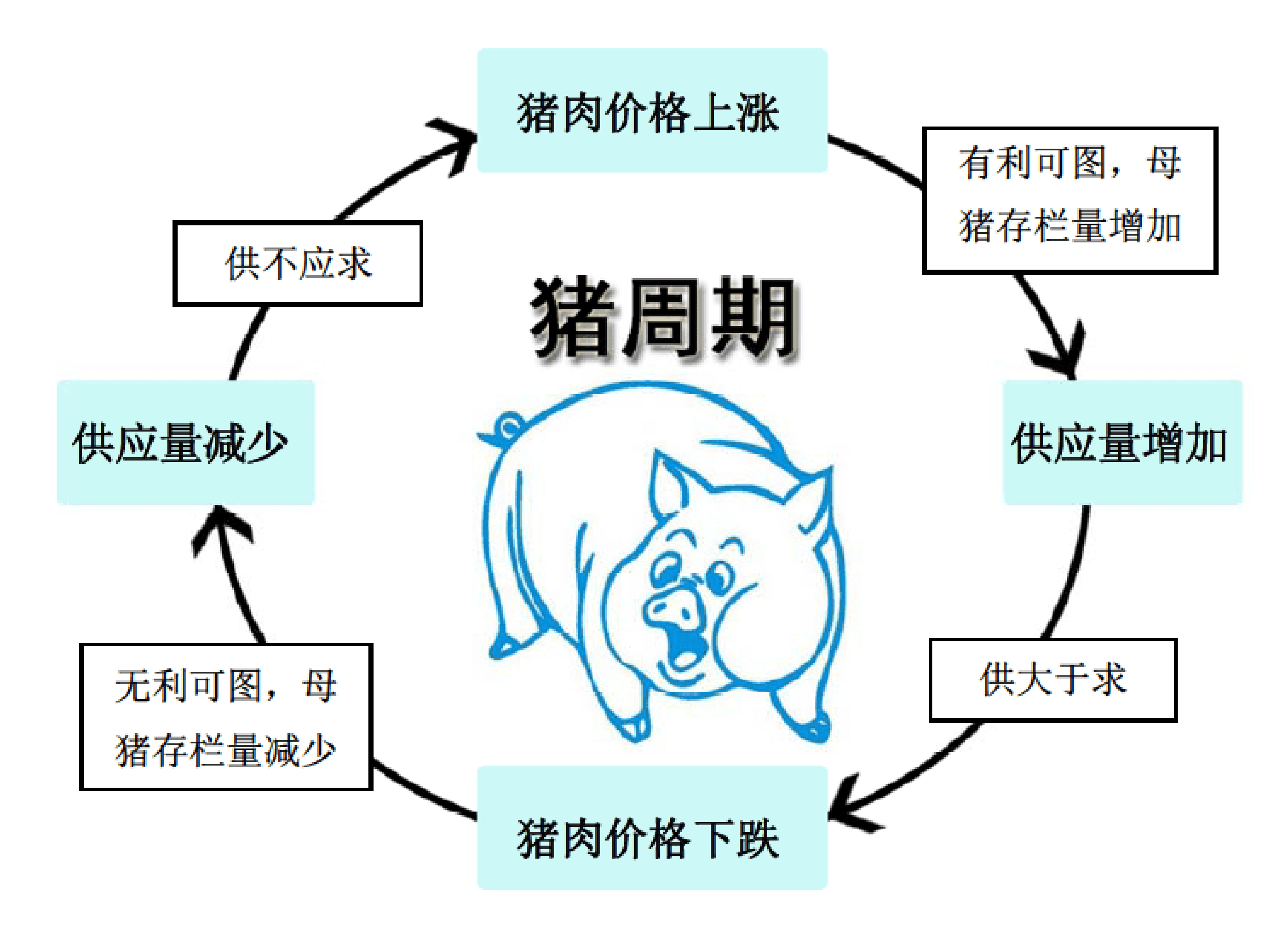


图 11-5 “猪周期”难题

1. 新希望结合大数据和云计算

新希望集团是中国农业产业化国家级重点龙头企业，中国最大的饲料生产企业和农牧企业之一，其拥有中国最大的农牧产业集群。

针对“猪周期”难题，新希望集团通过把历年的数据集中起来，建立一个动态的养殖、生产和市场的体系。通过大数据和云计算进行猪周期的预测，发现猪的价格波动周期有一定的规律，大概 3～5 年是一个完整的周期，少的时候两年多，多的时候 5 年多，而这个周期又受国家政策变化、天气变化、传染病变化、农民收入变化、原料价格变化等多重因素影响，同时又和人们的生活水准和购买力有关系。

新希望集团刘永好表示，如果全国所有养猪的农户都通过云计算、大数据对庞大的数据进行研究、分析、判断，研究出一个模型，建立信息系统，养猪就会变得更加科学化。

2. 温氏集团构建信息中心

广东温氏食品集团有限公司通过“企业＋农户＋客户”这一既分散又集约的生产模式，将分布在全国的 8000 多农户“化为一体”，户均年出栏生猪 800 多头，企业年出栏生猪约 680 万头，占全国年生猪出栏数的 1%。

“企业＋农户＋客户”模式是指，企业向农户提供猪仔、饲料及防疫技术，并负责市场销售，农户只承担养殖风险，无论市场周期如何变化，农户获利始终保持稳定，这在一定程度上化解了“猪贵伤民，猪贱伤农”的难题。另外，这种模式既有利于帮助农民就地实现就业，又避免大规模集中养猪的土地和环保压力。

温氏集团对生猪生产的安全控制非常重视，采用种猪统一培育、饲料统一生产、药品统一调配的全产业链条控制，确保了生猪的安全健康。农户管理员隔三差五就会上门

对农户进行指导和监测，他们利用 PDA 移动监控系统，在农户猪场就可以实现现场信息采集并实时传输到集团研究院数据中心进行运算分析，并实时提供解决方案。最后，在生猪上市前，企业会对每个农户的每批猪群都进行尿检，合格之后才卖给生猪批发商和肉联厂。

在温氏集团的研究院数据中心，电脑上可以清晰显示出生猪出栏价格的波动曲线，管理者可以实时监控全国 8000 多户养户的生产和出栏情况。

温氏集团的数据管理带来了很大的成效。2008 年猪肉价格步入下跌区间，广东新兴县城郊的温氏签约养户黄植强的养殖规模却持续扩大，他说："温氏的收购价格基本上没有大的波动，我的猪场增收也很稳定"。

【案例解析】：在本案例中，如何破解生猪生产大起大落的"猪周期"，走出"肉贵伤民，猪贱伤农"的怪圈，是道待解难题。笔者从业内人士处获悉，除了建立预警信息制度外，鼓励规模化养殖是解决"猪周期"的根本。散农户追涨杀跌的心理很强烈，这对市场的良性发展不利，而规模化养殖企业能主动获取市场信息，规避市场风险。规模化企业占得比重越大，养猪行业的组织化程度越高，生产才能有计划，价格也才能平稳可控，从而降低风险。

笔者认为，农业大数据其实还可以渗透到耕地、播种、施肥、杀虫、收割、存储、育种、销售等各环节，是跨行业、跨专业、跨业务的数据分析与挖掘。

11.2.6 【案例】钢铁企业用大数据摆脱困境

在"十一五"期间，济钢集团全力推进精准、高效、和谐发展战略，使决策更加科学，管理更加精准，运营更加高效，资源利用更加充分，努力成为中国一流、世界知名的现代化钢铁企业。

济钢集团作为老牌钢铁企业，在同行业中排名领先，与其雄厚的实力相匹配的是信息化建设的完善。济钢的信息化建设经过十多年的发展，已经拥有了基础自动化（L1）、过程自动化（L2）、产线管控（MES）、经营管理（ERP）、决策支持（BI）等信息系统，建立了完善的冶金信息自动化五级体系架构。在 2008 年之前，依托完善的冶金信息自动化五级体系架构，BI 系统的应用对济钢并不急迫。

不过，在 2008 年 9 月份，钢铁业面临经济危机的极大威胁，原料采购与钢材销售两大市场不可控因素越来越多，需要精细化管理提高管理效率，对企业的信息化建设提出了新的挑战。主要原因是由于 ERP、MES、计质量系统等内部运营信息系统有大量业务的历史数据的积累与沉淀，急需有效地从大量信息中提取有价值的分析数据和预测信息，来支持企业发展战略决策的制定。另外，随着市场形势的日益复杂，济钢集团建立一个数据来源于各业务系统、能整合外部数据并具高度可扩展性的决策辅助支持平台已迫在眉睫。

为了应对这一系列问题，2011 年 4 月 7 日，济钢集团成功实施了 IBM 的 Cognos 商业智能解决方案，帮助其提升企业内部的数据管理效率。商业智能业务分析，作为济钢管理信息系统完善提升项目的重要组成部分，已经成为济钢的核心应用系统。通过 Cognos 商业智能项目的实施，济钢集团的精细化管理得到了有效提升，决策更加准确，成本降低达到 20%以上。

专 家 提 醒

Cognos 在 BI 核心平台之上，以服务为导向进行架构，是唯一可以通过单一产品和在单一可靠架构上提供完整业务智能的解决方案。它可以提供无缝密合的报表、分析、记分卡、仪表盘等解决方案，通过提供所有的系统和资料资源，来简化公司各员工处理资讯的方法。作为一个全面、灵活的产品，Cognos 业务智能解决方案可以容易地整合到现有的多系统和多数据源架构中。2013 年 6 月 11 日，IBM 发布了 Cognos BI 最新版本 10.2.1，无论是重新设计的 UI 界面，还是新特性的加强，都给人耳目一新的感觉，如图 11-6 所示。



图 11-6 Cognos BI 10.2.1 界面

随后，济钢集团针对钢铁行业市场的严峻形势，在多个信息系统并行在建的情况下，以原料采购与钢材销售为切入点，进行 BI 系统调研、方案设计、系统实施，并于 2009 年 12 月 1 日上线 IBM Cognos 8，该系统将市场行情和内部运营情况清晰、直观地展现在公司领导面前，为公司快速应对市场变化、调整内部经营策略提供信息化支持手段。

以原料采购为例，2009 年时，与济南钢铁有长期协议的进口铁矿石出现货源危机，因此主要依靠购买现货来维持生产。但是现货的价格波动大，需要长时间地监控相关数据，预测价格趋势，并且要关注进口铁矿海运费的价格趋势，BI 系统的上线使这些问题迎刃而解。

在钢材销售方面，济南钢铁主要通过 BI 系统分析监测年度钢材销额、区域分布、大类情况、客户排名等，有效地掌握销售区域流向与客户排名，帮助企业快速调整销售策略。

2010 年，BI 系统帮助济南钢铁抓住三次商机，创造了两个亿的效益。其中市场行情分析室直接提报的分析报告创造价值 9000 万元，与采购部门合作赢得了 1.1 亿元。BI 系统对于行情分析室需要的上下贯通的分析，提供了有力的支撑。

【案例解析】：在本案例中，Cognos 系统的上线运行，为济钢集团掌控市场信息，科学合理采购、销售，更好地把握商机、降低成本、提高效益提供了重要帮助。

Cognos 展现的报表基于统一的元数据模型，统一的元数据模型为应用提供了统一、一致的视图。用户可以在浏览器中自定义报表，格式灵活，元素丰富，而且可以通过 Query Studio 进行即席的开放式查询。Cognos 有强大的报表制作和展示功能，利用它能够制作和展示任何形式的报表，其纯粹的 Web 界面使用方式又使得部署成本和管理成本降到最低。同时 Cognos 还可以同数据挖掘工具、统计分析工具配合使用，增强决策分析功能。

专家提醒

Cognos 具有独特的穿透钻取（Drill Through）、切片（slice）和切块（dice）、以及旋转（pivot）等功能，使分析人员、管理人员或执行人员能够从多角度对信息进行快速、一致、交互地存取，从而获得对数据的更深入了解，有效地将各种相关的信息关联起来，使用户在分析汇总数据的同时能够深入到自己感兴趣的数据细节中，以便更全面地了解情况，做出正确决策。

11.2.7 【案例】大数据提高企业核心竞争力

山东德棉股份有限公司始建于 1958 年，公司现拥有环锭纺 23 万枚、气流纺 3000 头，引进无梭织机 1214 台，是国家大型一档棉纺织企业。

2001 年以来，德棉的信息化管理逐步覆盖了财务、进销存、人力资源、生产管理、进出口业务等。但是，这些系统提供的数据完全面向业务，不能够满足领导的实际需求，成为科学、精确决策的瓶颈。德棉的 IT 投入像“正三角形”一样，大量的资金投入到了底层的基础架构建设方面，越向上越少。

总的来说，德棉集团的信息化存在三大问题，如表 11-2 所示。

表 11-2　德棉集团的信息化存在的三大问题

三大问题	具体对象	具体问题
角色定位	管理决策层领导在信息化系统中的角色定位问题	信息化究竟与领导的日常工作有什么关系？领导在信息化应用体系中处于什么样的位置，该扮演什么样的角色

续表

三大问题	具体对象	具体问题
知识壁垒	企业内部不同管理、业务的专业化导致的知识壁垒问题	由于各自所属的专业领域不同，在企业内部，业务部门普遍认为信息化就是信息技术部门的事情，而信息技术部门又无法以业务部门理解的语言表达信息化。此外，在业务部门之间以及业务部门与管理决策者之间也存在沟通的障碍
转化问题	实现从数据到信息，从信息到知识的转化问题	实现信息的业务化、管理化，首先要解决的是从数据到信息，从信息到知识的转化问题，而这一问题的解决则依赖于技术、产品支持

经过不断的摸索考察以及慎重的选择，德棉集团准备借助浪潮 ERP-BI 决策智能系统来解决这些问题。ERP-BI 决策智能系统的技术构架如图 11-7 所示。

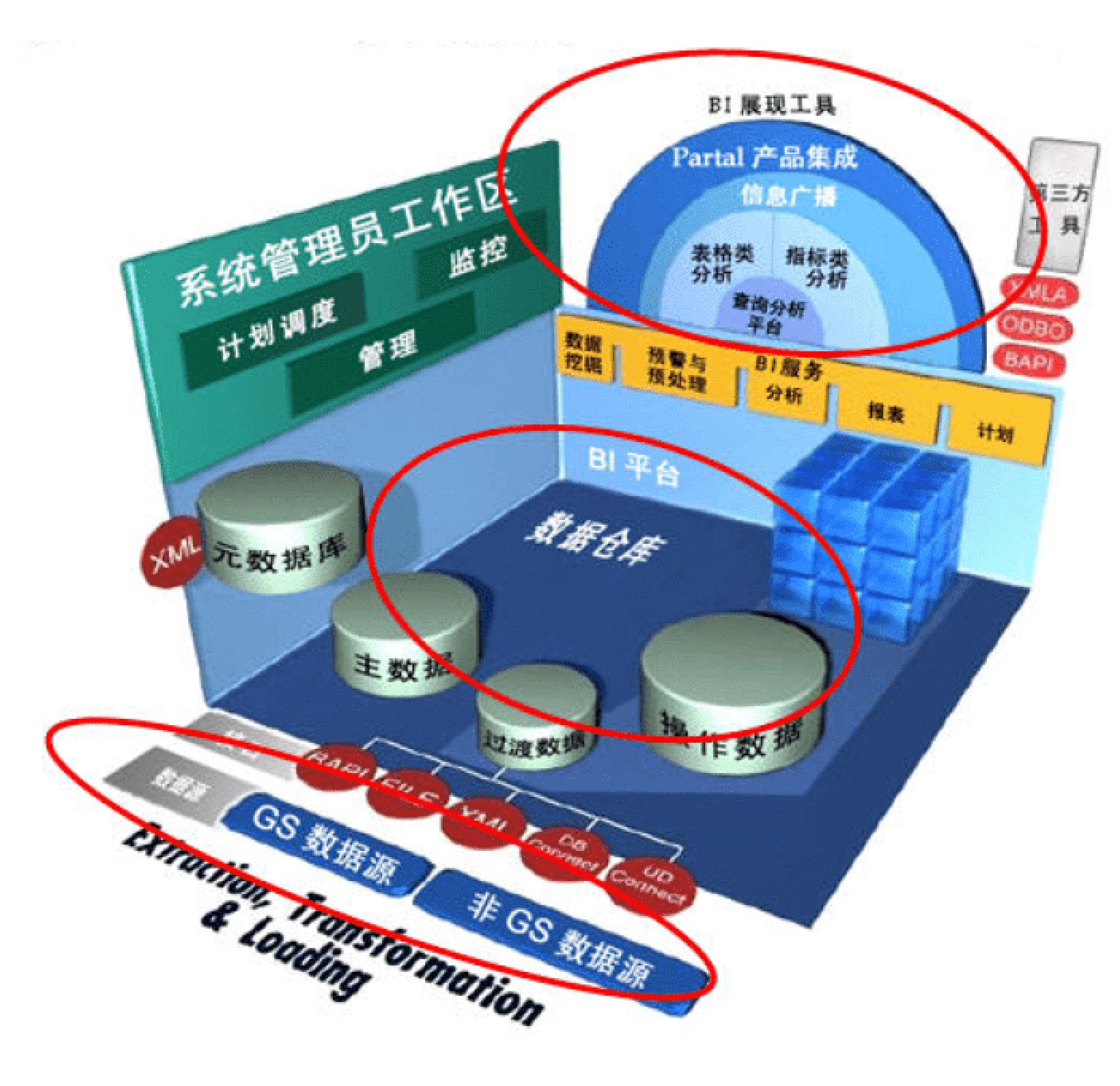


图 11-7　浪潮 ERP-BI 技术构架

专家提醒

浪潮 ERP-BI 系统包含系统模块、报表管理、万能查询、指标分析、管理驾驶舱、领导查询。报表管理模块管理集团公司的统一报表格式，实现在报表系统中查询各个子公司的数据，并且能够实现数据从报表到凭证的联查。万能查询将集团下发的标准格式导入到浪潮 ERP-BI 系统中，统计各种数据，能够正常查询科目明细账、凭证、增长排行、经营活动产生的现金流量表等，并且可以实现多表联查，能够立体展现集团公司及各个子公司的经营情况。

经过一段时间的紧张实施，德棉决策支持系统正式上线运行。随着单位内部信息系统的不断扩展，越来越多的数据被积累起来。这些数据包括：各种财务软件中的核算数据，分散在各地计算机中的 Excel、DBF 数据，分布在单位各类业务系统中的数据，各下级单位中的大量数据等。

功能强大与日益复杂的信息系统为单位带来了更多与更强的管理手段和方法，使德棉集团可以更好地规范管理，提高效率，确保管理的满意度。

该系统通过业务数据归集分析，抽取出集团领导关注的核心指标，做更高层次的抽象，把集团领导的精力从大量的表格、查询中解放出来，真正做到了让领导“享受技术，驾驭技术”，决策支持初现成效。

现在德棉各单位实施应用的软件有 11 个子系统、47 个主要模块、93 个子功能，分布在集团本部、各子公司，涉及业务包括财务、供应、销售、库存、进出口业务、生产计划、工艺、质量、统计、人事、工资、办公自动化、互联网应用、电子邮局、决策支持等。

【案例解析】：在本案例中可以发现，一个企业的信息分布在不同的部门和分支机构，决策者要综观全局、运筹帷幄，必须迅速地找到能反映真实情况的当前或历史的数据，并有效地预测未来；管理者要从不同的角度来审视和管理业务，必须从纷繁复杂的系统数据中迅速地找到数据与数据之间的关系，并获得各种统计结果和分析资料。这些正是德棉集团借助浪潮 ERP-BI 决策智能系统达到的目的。

笔者认为，企业要根据发展战略确定自己要解决的问题，利用数据分析找准发展瓶颈，在此基础上优化关键业务、核心流程的资源，从而通过信息化手段，提高自身的核心竞争能力。

读书笔记



12

餐饮：精准营销的数据

学前提示

衣食住行是人们的基本需求，所以很多人在创业时会把眼光放在这四大行业，而这其中属餐饮业竞争最为激烈。那么，餐饮经营者如何才能让自己的投资不打水漂，如何才能做到利润最大化呢？利用大数据精准营销的特点，即可帮助餐饮经营者通向成功。

要点点展示

- 餐饮行业大数据解决方案
- 餐饮行业大数据应用案例

12.1 餐饮行业大数据解决方案

餐饮业（catering）是将即时加工制作、商业销售和服务性劳动集于一体，向消费者专门提供各种酒水、食品、消费场所和设施的食品生产经营行业。餐饮市场整体上供大于求，处于过度竞争的状态，因此做好定位至关重要。面对着这个市场信息爆炸的时代，餐饮业数据挖掘该怎么做，要如何利用大数据进行准确精准的餐饮市场定位呢？本节将重点分析餐饮业数据挖掘的市场现状和前景。

12.1.1 大数据在餐饮业的市场现状

俗话说："民以食为天。"长期以来，餐饮业作为第三产业中的主要行业之一，对刺激消费需求，推动经济增长发挥了重要作用；在扩大内需、安置就业、繁荣市场以及提高人民生活质量等方面，都做出了积极贡献。

随着我国居民消费水平的快速提高，人们追求品牌店、特色店和名牌餐饮店的势头更加明显，个性化特色经营突出的品牌、特色餐饮深受青睐。中国餐饮业的发展趋势如表 12-1 所示。因为看到行业前景和利益驱动的原因，进入这一领域的经营者必然会大大增加，不可避免地要带来激烈而残酷的竞争。

表 12-1 中国餐饮业的发展趋势

发展趋势	具体表现
个性化消费日趋明显，特色餐饮更趋突出	市场消费从以价格选择为主向价格、品位、氛围、服务和品牌文化等综合方向发展，注重选择的理性化消费特点增强，个性化和特色化成为广大消费者和企业经营共同追求的时尚。为满足个性化需要，要求企业不断提高经营的特色与水平
连锁经营迅速发展，企业发展多元化趋势增强	以连锁经营为代表的现代餐饮业加速替代传统餐饮业手工随意性生产、单店作坊式经营、人为经验型管理，向产业化、连锁化、集团化和现代化的方向迈进。餐饮业所有制结构已发生了根本性的变化，在行业规模企业发展中，投资主体多元化、经营模式多样化和企业规模化、集团化趋势日益明显，实力逐步增强
安全、健康、卫生的餐饮场所成为消费者的首选	受地沟油、禽流感等的冲击，餐饮市场从传统的色香味型，并以味为主转为更加注重安全卫生、健康营养的消费。安全、健康的餐饮消费成为餐饮企业与消费者的共同追求，餐饮企业经营者行为规范，促进了餐饮企业质量的提高

这样的大背景对餐饮经营者的决策产生了更高的要求。面对全行业过度竞争的局面，如何创造局部的优势，对全体餐饮人来说是很大的挑战。如果在一个细分市场没有

优势，就会陷入到同质化的竞争中去，这对企业的生存和发展都将是非常不利的。这些优势有可能是局部的优势，有可能是地点或地域的优势，也有可能是一部分特征人群的优势。

因此，餐饮企业的目标应该是在不同的细分市场创造局部优势，如此就能在一个完全竞争的环境中，赢得相对的垄断地位，为企业带来生存上的保障。例如，夜宵诱惑的核心顾客应该是加班族，针对主要顾客层级，企业要从选址、产品、服务价格等一系列环节进行调整，当然前提是需要依靠数据的准确采集与提供。

12.1.2 餐饮行业面临的大数据挑战

中国菜也是世界上最全面、最丰富的菜别。可是为什么中国餐饮一直做不大呢？面对外国餐饮企业社会化生产和规模化经营，依靠经验型管理和传统式经营的中国餐饮企业，显然处于劣势。尤其是在大数据时代，我国餐饮行业将面临以下三大挑战。

1. 如何控制餐饮成本

目前，餐饮行业的竞争环境发生了很大的变化，主要是三类成本上升迅速，如表 12-2 所示。

表 12-2 餐饮行业的三类成本

三 类 成 本	细 节 内 容
人力成本	人事费用包括了员工的薪资、奖金、食宿、培训和福利等
原材料成本	是指餐饮成品中具体的材料费，包括食物成本和饮料成本，这也是餐饮业务中最主要的支出
经营成本	包括租金、水电费、设备装潢的折旧、利息、税金、保险和其他杂费

近期餐饮行业面临更大的压力，体现为原材料成本、房租成本的迅速提高，利润率下滑是目前餐饮行业基本的状态。人力成本和房租成本的上升是必然趋势。在大数据时代，如何控制成本成了餐饮行业首要解决的问题。

2. 如何进行多渠道消费

在社会消费的引领者中，多渠道消费的特点十分明显。目前，大部分餐饮企业都是采用实体店经营的方式，在多渠道消费上的注意力则略显不足。如何在多渠道消费领域升级服务，是摆在餐饮行业每一个企业家面前的难题。

现阶段，一项全新的信息化应用服务——餐前的网络订餐悄然兴起，有的企业自建了订餐平台，有的则使用第三方服务平台为消费者提供网络支付和商家结算。顾名思义，网上订餐就是互联网的深入应用，其流程如图 12-1 所示。用户通过互联网，能足不出户，轻松闲逸地自己选购餐饮和食品（包括饭、菜、盒饭、便当等）。随着食天下网上订餐平台的兴起，网上订餐已经逐渐成为了白领阶层中的一种潮流了。



图 12-1　网上订餐的流程

到餐厅以后的定位点菜，实现的主要工具是平板电脑、智能手机等，客户在用餐过程中可以进行抽奖活动，用餐之后还可以利用点评网进行点评。现阶段消费者已经越来越倾向于多种渠道的消费模式。

3. 如何跟上大数据的步伐

由于历史和技术的种种原因，餐饮企业的信息化建设缺乏长期的规划，因此逐渐形成了信息孤岛。当企业规模达到一定程度时，这些孤岛便成为影响企业运营效率和流程的阻碍，由此财务流程一体化的协同管理将成为未来的主流应用。

云计算、新媒体等新技术的快速发展，成为推动社会发展的重要因素，其对餐饮业的影响也很深远，这些新应用正潜移默化地改变着餐饮行业的发展方式。

如今，很多餐饮企业都转而应用云计算，摒弃了原来繁杂的 CS（Customer Satisfaction）顾客管理方式。运用云计算可以有效降低管理成本，快速升级、快速部署，更为迅速地对市场和消费者需求进行反应。现在很多餐饮企业大幅度增加 IT 方面的投资，强化信息化技术管理，加速推动整个餐饮行业的 IT 信息技术建设。

另外，很多餐饮企业已经由原来的业务管理信息化，逐步提升到业务管理精细化。很多具有一定规模的餐饮企业已经完成了核心业务模块的信息化建设，接下来的重点是管理信息化向管理精细化的过渡，向管理要效益。

在餐饮行业，数据的挖掘和分析也将得到更多的应用。餐饮行业已经积累了大量的历史数据，如何有效地利用这些数据，需要专业的工具和手段支持。企业对用户就餐体验的深入关注，将会使智能终端的应用越来越广泛。

为了给消费者提供更好的消费体验，餐饮企业内的智能终端应用将越来越丰富，目前两种主要的终端应用平台一个是基于 iOS 系统，另一个是基于安卓系统。餐饮企业通过设备和顾客进行深层的互动，获得消费者的评价信息和调查数据。

4. 餐饮企业如何面对数据的挑战

目前，餐饮行业很少提及大数据这个概念，毕竟中国信息化建设只有 30 年左右的

时间，具体到餐饮业充其量不超过 10 年。因此，餐饮行业信息化的建设仍然属于“人治”的状态，随意性比较大，尚未形成信息化和规范化的管理制度，缺乏对信息化的实施和控制。信息化决策机制不完善，风险管理缺位，数据没有使用起来，导致企业管理很大程度上要依赖于个人领导力，这也会增长信息化的风险和不确定性。此外，餐饮行业也存在找不到信息化中心的问题，这些都会影响信息化的成功实施。

对于还未施行信息化策略的中小餐饮企业来说，首要任务是使用信息技术来提高自身的管理水平，把中国的传统饮食与现代信息化管理有机地结合在一起，为企业的做大、做强、管理规范化提供支撑。餐饮企业的管理目的是成本控制、运营控制，其最终结果表现为效率和效益。而要达到这一目的，管理数据的及时性、准确性、完整性、有效性是至关重要的，而这些特性恰恰是信息系统最重要的特性。

对于已经做好 IT 规划的成长型餐饮企业来说，要有一定的前瞻性，制定三五年的中长期规划，避免信息规划不统一，甚至产生信息孤岛的情况。信息规划是动态匹配的过程，是用具体的 IT 技术最大程度地解决和满足企业的业务需求的过程，所以在 IT 规划前必须先进行组织业务的规划。

12.1.3 大数据对餐饮企业有何作用

在大数据时代，集成化和个性化是企业运营的典型特征：

（1）集成化。系统的集成直接产生了“小前台（智能终端）+大后台（大数据）”的经营模式，切实简化了前台的操作。这也符合餐饮行业的整体趋势，前端的简化将有效减少系统使用的培训工作。集成化另一种方式是数据的集成，例如，银行在这方面先行一步，当我们外出消费时，通常会收到银行的短信提醒，内容是消费金额，以及获得什么样的积分奖励等内容。

专家提醒

云计算可以说是集成化模式下的典型应用，这种应用操作成本比较低，必将成为主流，目前应用主要集中在网络点餐等方面。

（2）个性化。数据的个性化是通过集成化来实现的，在识别出顾客的个性需求后，企业就可以针对顾客进行个性化服务。如何才能使数据的挖掘和应用做得更好？需要企业对消费行为有更深刻的识别。

专家提醒

笔者认为，除了系统和数据层面的集成化，电子商务也得到了快速发展。目前，很多餐饮企业都在进行电子商务方面的尝试和探索，例如和团购网站合作，使用第三方平台的订餐系统，也可以自己搭建 B2C、B2B 平台。

数据时代信息化的作用，还可以延伸到品牌的宣传中：

（1）传播途径变广。现在的传播形式已经发生了显著变化，对微博、互联网、微信、二维码的应用是餐饮企业在未来发展中的必经之路。

（2）更快地提高效率。信息化可以减少繁琐的手工操作、员工数量和工作复杂程度。

（3）提高整个团队的管理水平。在财务供应链的信息化建设过程中，通常伴随着流程的改变，因此通过信息化可以固化和优化流程，从而达到提升组织管理水平的目的。

12.1.4　餐饮企业该如何应用大数据

经营餐饮业需要相当高明的营销艺术，将最好的构想变为噱头，尽量做到“人无我有，人有我精。”只有以客人为中心，以市场为导向，改变经营观念，才可以处于不败之地。

因此，在餐饮行业中，大数据不能大而无用，要对应到特定企业、特定人群、特定需求上，才能发挥特定作用，产生价值。针对餐饮企业特定需求的数据支撑服务，针对特定人群的特定需求的数据支撑服务，就是大数据的“小而美战略”。做创新的餐厅项目，要记住小而美、少而精的细分领域，主题餐厅结合特定目标群，设计品种丰富但单品少而精。

下面是大数据在餐饮企业的具体应用，如表 12-3 所示。

表 12-3　大数据在餐饮企业的具体应用

企业应用	具体内容
基于 LBS 的地理位置服务	LBS 服务可以用来辨认一个人或物的位置，例如发现最近的提款机或朋友同事目前的位置，也能根据客户目前所在的位置提供直接的手机广告，包括个人化的天气信息提供，甚至提供本地化的游戏。现在消费者需要餐厅位置信息的相关服务，而现有的服务商并不能完全理解消费者的意图，也不了解客户知道这些信息后的行为，何种服务才能吸引用户。因此，能够提供实时信号、地理位置、在线活动和社交媒体，并支持众多其他类似情景的综合服务，将是今后的趋势与主流
企业数据在管理决策中的应用	通过 SCM 管理系统，可以对采购价格进行分析，生成采购价值指数，对数量、价格这些因素进行全面、系统的分析；同时通过 CRM 系统，对顾客的消费行为进行更深层次的挖掘与分析
企业基础数据管理	运用大数据系统可以管理酒菜设置、特价促销、酒菜折扣、酒菜组成、房台设置、消费方式、员工资料等
规避经营风险	运用大数据系统可以充分洞察和分析餐饮管理的现状，并对企业管理的流程有深刻的理解和准确的把握，帮助企业利用计算机强大的数据处理能力和流程优化能力，实现自动化管理，简化企业的工作流程，减少浪费及人为管理的疏漏现象，重新优化配置企业资源，把经营成本降到最低

12.2　餐饮行业大数据应用案例

大数据技术的发展，将餐饮业的竞争带入了一个全新的境界。正当的竞争给了餐饮业的发展无穷的动力。那么，大数据的日渐普及又给餐饮业带来什么机遇和挑战呢？笔者认为，大数据技术除了带给餐饮企业与顾客交流沟通的高效、便捷外，最大的好处便是可以通过餐饮管理软件和网站来建立自己的客户数据库。对于餐饮企业，特别是大规模的连锁餐饮企业，拥有自己的客户数据库，无疑于在信息时代占领了市场竞争的战略制高点。本节主要介绍餐饮行业大数据的应用案例，希望对读者有一定的启发和学习价值。

12.2.1　【案例】农夫山泉用大数据卖矿泉水

在上海某个超市的一个角落，农夫山泉的矿泉水静静地摆放在这里。来自农夫山泉的业务员每天例行公事地来到这个点，拍摄 10 张照片：水怎么摆放、位置有什么变化、高度如何……

这样的商店每个业务员一天要跑 15 个，按照规定，下班之前 150 张照片就被传回了杭州总部。每个业务员，每天会产生的数据量在 10MB，这似乎并不是个大数字。不过，把范围再扩大一点，这个数据就会变大。农夫山泉在全国有 10000 个业务员，这样每天的数据就是 100GB，每月为 3TB。

挖掘这些数据到底有什么用呢？农夫山泉面对这些照片，很快找到了几个突破口。怎样摆放矿泉水更能促进销售？什么年龄的消费者在水堆前停留更久，他们一次购买的量多大？气温的变化让购买行为发生了哪些改变？竞争对手的新包装对销售产生了怎样的影响？农夫山泉从 2008 年就开始收集这些照片，如果按照数据的属性来分类，“图片”属于典型的非关系型数据，还包括视频、音频等。要系统地对非关系型数据进行分析是农夫山泉在“大数据时代”必须迈出的步骤。

1. 营销信息化方案

2007 年底，农夫山泉决定甩开经销商，自己控制营销市场，并着手建立一支直接面向终端的一线业务代表团队。农夫山泉将工作的重点转向了营销信息化，开发了营销管理短信平台，借助 GPS 服务和全球定位增值业务，把每一个经销商、终端门店和终端业务员的销售数据都集中起来管理。

借助手机终端，农夫山泉实现了对业务代表和销售人员的实时监控、管理，公司的管理触角直接由一级经销商扩展到零售门店，甚至直达终端消费者，从而牢牢掌握住了渠道。而以电子数据流作为依据，从订单到收货，农夫山泉也能够随时查询、分析所有

的数据信息，为决策提供支持。

目前，除了中国香港和台湾地区，国内所有省市都有农夫山泉的业务员在使用手机终端运作业务，每月总短信量高达 1000 万条之多，覆盖范围极广。

2. 携手 SAP 大数据

早在 2004 年，农夫山泉就引进了 SAP 的 ERP 系统，不过当时的农夫山泉仅仅是一个软件的采购和使用者，而 SAP 也还只是服务商的角色，因此效果并不理想。2011 年 6 月，SAP 和农夫山泉开始共同开发基于“饮用水”的产业形态中，运输环境的数据场景。

农夫山泉在全国有十多个水源地，通过把水灌装、配送、上架，一瓶超市售价 2 元的 550ml 饮用水，其中就有 3 毛钱花在了运输上。因此，如何根据不同的变量因素来控制自己的物流成本，成为农夫山泉的核心问题。

在采购、仓储、配送这条线上，农夫山泉特别希望大数据获取解决三个顽症：首先是解决生产和销售的不平衡，准确获知该生产多少，送多少；其次，让 400 家办事处、30 个配送中心能够纳入到体系中来，形成一个动态网状结构，而非简单的树状结构；最后，让退货、残次等问题与生产基地能够实时连接起来。

对此，SAP 团队和农夫山泉团队开始了场景开发，他们将很多数据纳入了进来：高速公路的收费、道路等级、天气、配送中心辐射半径、季节性变化、不同市场的售价、不同渠道的费用、各地的人力成本甚至突发性的需求（例如某城市召开一次大型运动会）等。

2011 年，SAP 推出了创新性的数据库平台 SAPHANA，农夫山泉则成为全球第三个、亚洲第一个上线该系统的企业，并在当年 9 月宣布系统对接成功。采用 SAPHANA 后，同等数据量的计算速度从过去的 24 小时缩短到了 0.67 秒，几乎可以做到实时计算结果，这让很多不可能的事情变为了可能。

2013 年，农夫山泉再次携手 SAP，尝试开发基于 SAPHANA 的 SAP Business Suite。农夫山泉希望借助这一最先进的业务平台，在实时分析海量数据的基础上，加快应收应付账款管理、简化订单流程、优化库存管理、加速物料资源计划运算，从而在各种“端到端”业务流程中实现全新的商业价值。

有了强大的数据分析能力做支持后，近年来，农夫山泉以 30%～40%的年增长率，在饮用水方面快速超越了原先的三甲：娃哈哈、乐百氏和可口可乐。根据国家统计局公布的饮用水领域的市场份额数据，农夫山泉、康师傅、娃哈哈、可口可乐的冰露，分别为 34.8%、16.1%、14.3%、4.7%，农夫山泉几乎是另外三家之和，如图 12-2 所示。

【案例解析】：在本案例中，作为一家后来居上的快消品企业，农夫山泉的产品线并不像可口可乐、康师傅、娃哈哈那么齐全。在此背景下，它凭借与之同台竞技的资本就颇值得仔细推敲。除依靠特色产品之外，狠抓渠道管理、重视终端市场表现，并借助

IT 系统制定出快速反馈机制，是农夫山泉的秘密武器。

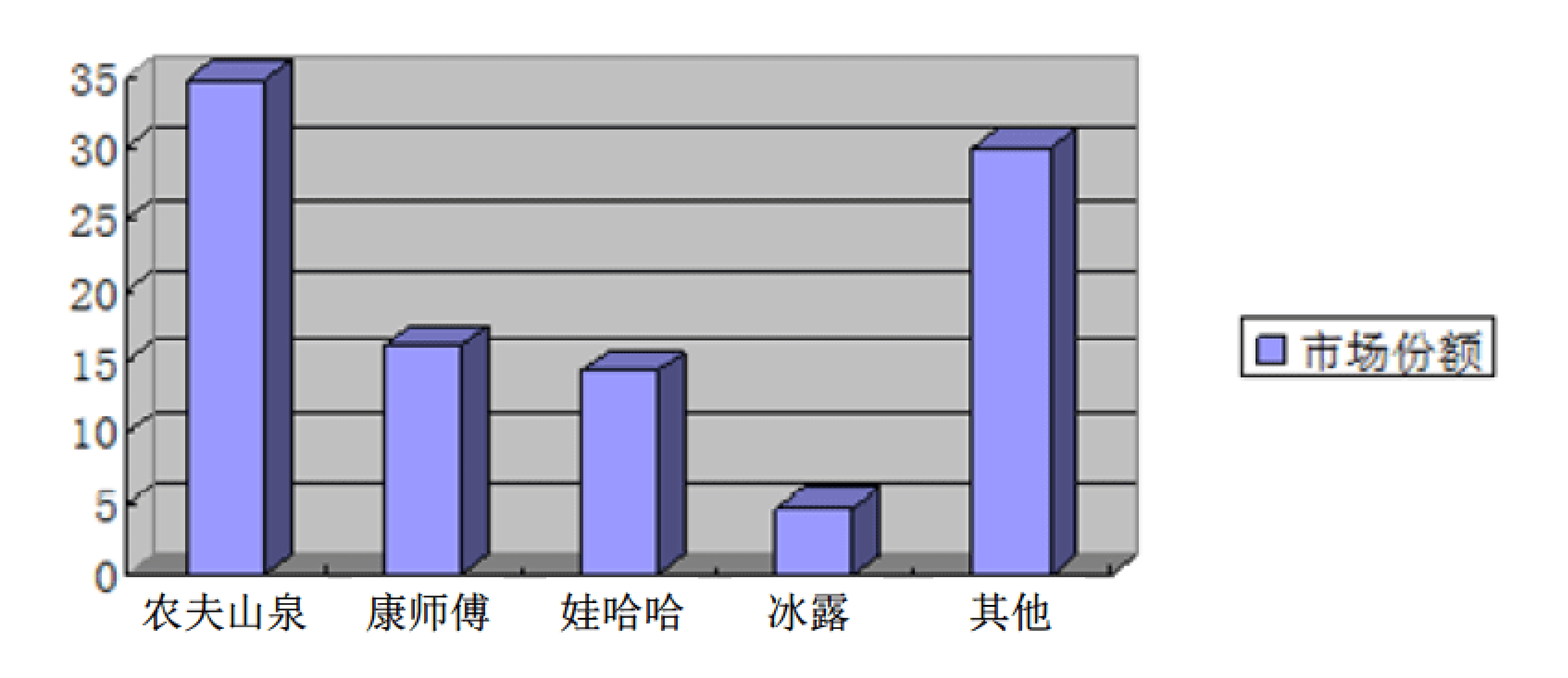


图 12-2　2012 年饮用水品牌市场份额

笔者认为，企业对于数据的挖掘使用可以分为以下三个阶段：

- 首先把数据变得透明，让大家看到数据，能够看到的数据会越来越多。
- 然后可以提问题，可以形成互动，很多支持的工具来帮助我们做出实时分析。
- 最后，通过信息流来指导物流和资金流，即用数据预测未来，指导企业前进的方向。

12.2.2　【案例】绝味鸭脖的大数据经营模式

“绝味”意为绝妙的、绝无仅有的味道，其经典美味的鸭脖深得消费者青睐。绝味全国门店现已突破 5000 家，从创办至今，共累计服务顾客达 10 亿人次，已成为鸭脖连锁领导品牌。

鸭脖的产业规模令人惊讶，2013 年达到了近 370 亿元的市场容量和规模，对于绝味来讲，每天约有 70 万人次走进绝味的门店，平均每天售出 100 万根鸭脖，2013 年的年零售额已经接近 40 亿。

一根看似毫不起眼的小鸭脖，能达到这样的规模，这是很多人没有想到的。目前，绝味已经为 3300 个加盟商实现了创业梦想，解决了 20000 名员工的就业问题。这一组数据揭示了绝味正是“小行业大市场”的企业典型，通过小小的鸭脖，绝味撬起了巨大的休闲熟食市场。

绝味在商业模式、管理方式、营销手段上都做了创新和尝试，才获得了今天的地位，主要表现在以下 3 个方面：

（1）销售模式的变革。绝味引入了特许经营这样一个商业模式完成了零售业态的改变。

专 家 提 醒

例如，随着微信公众平台的推出，作为行业领导品牌的绝味敏锐地把握了这一极具价值

的推广资源，已正式开通微信平台。使用微信的人大多年轻、时尚，追求新事物，这和绝味的目标人群相匹配。通过微信平台，绝味能实现对目标人群“点对点”的信息推送和实时互动，并保证高效到达。微信平台将成为绝味和消费者之间最快捷的数据沟通桥梁。通过这一新媒体，消费者可以更方便、及时地了解绝味的相关信息、资讯，更便捷地参与绝味推出的活动，享受到更多的优惠等。同时，绝味也可以提高消费者黏性，实现品牌的“病毒式传播”。

（2）管理方式的改变。绝味将传统的作坊式工厂、门店上升到规模化生产，同时实现了管理干部以及人才梯队的搭建。

（3）采用数据决策。绝味导入了信息化建设，专项资金接近两个亿。特别是导入了世界 500 强的先进管理工具 SAP，在传统食品制造行业尤其是在卤制食品制造业是第一家。

SAP 是目前全世界排名第一的 ERP 软件。根据应用场景的特性，SAP 针对性的数据库可以分为 5 种：行式数据库、列式数据库、内存数据库、嵌入式数据库、数据流处理等。由于客户数据的交易、迁移、存储、分离、分析都各有特点，之间不可能含糊，不可能都用一个技术解决所有问题。基于此，SAP 在各个细分市场上提供了相应的数据库产品：在分析型数据库方面，Sybase IQ 有最佳的 TCO 表现；在交易型数据方面，SAP 的 Sybase ASE 有最佳的 TCO；在移动以及嵌入式数据库方面，SAP 有 SQL Anywhere；在统一的实时数据管理平台上，SAP 也有对应的产品。

目前，绝味正在努力打造一流的特色美食平台，让“绝味”成为汇集各类美食的渠道，让消费者更便捷地获得健康、安全的美食。同时，绝味还将强势推广品牌，不断拓展经营加盟商，为加盟商提供更好的加盟环境，进而实现与消费者、加盟商三方共赢的商业生态圈。

【案例解析】：在本案例中，绝味通过与消费者的近距离深入互动，进一步融入了消费者的休闲生活，这不但提高了绝味品牌的美誉度，还是其进入互动营销新时代的标志。

笔者认为，绝味还可以让大数据飞得更高，它是雄才大略者的利器，它将使企业具有无可比拟的竞争优势。

12.2.3 【案例】“哆啦宝”打造精准营销平台

2013 年 7 月，国内第三方支付企业易宝支付正在低调测试一款餐饮营销类产品“哆啦宝”，欲凭借其掌握的支付数据反向尝试客户管理和精准营销。这也标志着易宝支付将从消费后端的支付环节正式涉足消费前端的营销环节。

“哆啦宝”是针对线下商户的智慧支付营销解决方案，是集硬件智慧营销终端 POS、软件会员营销解决方案、商户网络营销平台以及社会化媒体营销平台于一体的效果营销

解决方案，掀起线下支付营销按效果付费的风潮，帮助企业一起挖掘“消费后市场数据”，如图 12-3 所示。



图 12-3　“哆啦宝”的营销特点

“哆啦宝”主要面向餐饮类商户，它在商户 POS 机中内置一套系统，该系统主要用于采集用户交易数据、进行客户管理。消费者第一次在商家刷卡消费时，在 POS 机上输入手机号，可以短信收到商家的红包信息，同时将手机号与其银行卡绑定。下次到店刷卡消费时，POS 机内置系统将自动识别红包信息，并扣掉相应优惠金额，并再生成一个红包，以此循环。

数据显示，2012 年，我国银行卡消费金额达 20.8 万亿元，共 90.09 亿笔，分别较 2011 年增长 36.9%和 40.5%，随着银行卡的普及，刷卡消费额更是同比增长超过 50%，占社会消费品零售总额的比重超过 40%。

易宝支付也发现了其中的大机遇，并与近百家金融机构达成战略合作关系，并支持 34 家银行卡升级为红包银行卡，普通银行卡只要刷“哆啦宝”POS，即可为此卡创建一个红包账户，完成智慧升级。此后，在任何一家“哆啦宝”合作商户刷卡，即可获得消费后商家返的现金红包，并直接存入红包银行卡，下次刷卡消费可自动抵现。此模式不改变商户使用传统 POS 的任何操作，无声无息地帮商户完成消费后营销，同时消费者只需激活一次，即可尽享“哆啦宝”合作商户的个性化优惠折扣。

另外，通过内置的软件，POS 机产生的每一笔刷卡交易都将在“哆啦宝”形成记录，“哆啦宝”则将基于交易数据做精准的客户营销，提高二次消费率，其营收则主要通过商户返点获得。据悉，“哆啦宝”试运营期间，合作商户的回头率已高于 16%，超过了团购行业大约 10%的回头率。据悉，易宝支付旗下目前在全国铺设了约 10 万台 POS 机商户（占全国终端 POS 机总数的 1.7%左右）。

【案例解析】：以餐饮行业为主的生活市场已经盘踞着形形色色的大型企业以及创

业公司，也总有全新的模式时常跳出来吸引市场的眼球。在本案例中，易宝支付的“哆啦宝”代表了一种近期正在流行的新趋势：深入商户后端，精细化运营老客户，而不是一味追求前端营销。

“哆啦宝”的服务归根结底也最有价值的部分其实是数据服务。笔者认为，从大数据角度来看，集体用户的行为规律很重要，但搞清楚用户是谁更重要。用户的行为规律可以作为改善老产品、生产新产品的有力依据，而知道用户是谁、在哪，并能随时随刻地找到他们、触及到他们才是完成商业转化的关键。

12.2.4 【案例】打造适合你的找餐馆手机 APP

“好友美食”APP 是基于新浪微博的开放社交图谱制作的，其通过提炼 6000 万微博数据，可以帮助用户通过社交好友发布的内容获得好友喜欢的美食。

“好友美食”会直接根据用户所处的地理位置，在首页向用户推荐附近的美食，并显示推荐理由、实际距离、人均消费等信息，如图 12-4 所示。点击进入每个店铺的单独页面，除地址、电话等基本信息外，“微博评价”以不同的冷暖色调呈现出来，暖色为正面评价，冷色为负面评价，黑色则为中性评价。

例如，你请朋友吃饭，去一家自己喜欢的川菜馆，结果朋友不喜欢吃辣。这就是典型的无法根据好友兴趣挑选美食的问题。“好友美食”通过抓取新浪微博大量关于美食的数据显示，仅含有“麻辣诱惑”的关键词就抓取了上百万条信息，可以覆盖数百万用户的个人喜好。再例如，某个用户的微博内容中包含一次“火宫殿”一词，也许不能说明什么，但是如果“火宫殿”这个词出现了 5 次以上，那么至少证明他经常去这家餐厅吃东西。

目前，“好友美食”的基础数据来自于新浪微博的开放平台，经过数据挖掘分析后呈现给用户，涵盖了北京、天津、武汉、杭州、西安、上海、成都、重庆、广州、深圳、南京等 11 个城市，未来会陆续加入其他城市。

除此之外，“好友美食”允许随时随地拍照上传到新浪微博，同时也会显示你的好友在附近哪家餐厅吃过的评价，你的好友也能同步看到你对于餐厅的评价，每位用户都会有自己对之前吃过的地方发布评价的记录，可以看到你自己还有好友的美食轨迹，为自己的吃货道路留下每一份记忆，如图 12-5 所示。

【案例解析】：在本案例中，对于餐饮企业来说，“好友美食”APP 对顾客意见信息的收集获取，对其经营发展有重大意义。餐饮企业可以对这些数据进行分析整理，找到企业在经营管理上存在的不足和缺陷，然后针对此进行有目的的调整和改善，只有这样，企业的经营管理水平才会得到不断地提升和进步，才会赢得更多顾客的喜爱和认同。





图 12-4　向用户推荐附近的美食

图 12-5　发布评价

笔者认为，餐饮企业还可以利用这些数据来分析顾客的饮食习惯，达到精准营销的目的。另外，还可以针对主力目标用户群的生活需求和精神需求，和一些品牌商家联合做沙龙或体验式活动，为用户提供他们需要的其他种类的产品。例如，对于女性用户的消费和购买需求，除了美食，还会有各种护肤美容、休闲健康、服装搭配等方面的需求。联合做活动，进行用户数据的收集、跟踪和汇总分析，挖掘数据的价值，形成数据服务，提供增值服务，创造商业价值。


读书笔记

13

金融：大数据理财时代

学前提示

金融事务需要搜集和处理大量数据，对这些数据进行分析，发现其数据模式及特征，然后可能发现某个客户、消费群体或组织的金融和商业兴趣，并可观察金融市场的变化趋势。本章将针对金融行业，探索大数据时代的企业理财经。

要点展示

- 金融行业大数据解决方案
- 金融行业大数据应用案例

13.1 金融行业大数据解决方案

互联网的发展和信息爆炸已经将我们推入了以云计算和大数据为新特征的信息社会，数据爆炸性增长催生了大数据技术的出现，引发了一系列衍生物出现，如互联网金融等。大数据已经不再只是实验室的研究课题，它们已经对社会造成了冲击，并对商业实践产生了颠覆性的影响。金融业作为传统行业之一，也感受到了“数据地震”，金融机构若不能紧随经济、技术和社会的发展而发展，就会面临被淘汰的危险。

13.1.1 大数据对传统金融行业的影响

从现代信息技术的潮流看，近两年来全世界掀起了一波大数据的浪潮，美国奥巴马政府宣布了“大数据的研究和发展计划”，欧盟也明确提出了“开放数据战略”。如何在大数据时代更好地推动金融创新，是传统金融行业必须认真面对和严肃思考的问题。

对金融行业来说，使用“大数据金融”的概念，制定并实施“大数据金融”战略，更能体现金融业自身的实力和潜力，也更能与网络业及其他行业有机融合，平等竞争，在大数据时代找到自身生存发展的机会也更大。

如今，世界正在步入大数据时代，为后来者提供了不可多得的战略空间和机会。例如，京东商城、金银岛等电子商务企业借助平台积淀的数据资产纷纷进军供应链金融领域，将信息流、物流和资金流深度融合，为平台上的用户提供订单融资、仓单融资等服务。该模式弥补了传统供应链金融信息技术支撑不够、服务范围有限等不足，推动了供应链金融的进一步发展。

在大数据时代，传统金融机构也开始采取积极的应对措施，以面对新兴金融力量的不断渗入造成的威胁。例如，银行业推出网上银行、网络融资和电子商务等业务，保险业亦开始探索通过网络销售保险、网上个性化保险产品和虚拟财产保险等业务。

然而，对于金融业这么一个数据密集型行业来说，无论是传统的线下业务还是新型的线上业务，数据仍然是其竞争的关键要素。银行业进军电子商务的核心目的在于采集数据，银行业开展网络融资、保险业探索虚拟财产保险的成败关键则在于利用数据。由此可见，大数据俨然成为金融业构建核心竞争力的重要资产。

对传统金融企业来说，是否以自己为中心提供各种网络服务已经变得没有过去那么重要，获取和利用他人所产生的数据变得更加重要。基于某种服务所积累的数据价值在贬值，数量再多也算不上大数据，只有获取网络世界中全面的数据才有深度整合利用的价值。正因如此，传统金融企业就大可不必邯郸学步，重复互联网运营商走过的道路，非要先建立各种非本业服务以获取本业之外的数据。

笔者认为，传统金融业在新的历史环境中面临机遇与挑战，因此，必须利用大数据的理念改造自身。抓住大数据的机会，是中国金融业新时代的使命所在，企业可以利用自身优势探索一条新路。

专家提醒

与其他传统产业相比，金融服务业是电子化、网络化和数据化程度最高的产业之一，也许仅次于网络和电信业。由长期的金融服务积累的数据完全可以在确保用户隐私和商业机密的前提下，与各行各业共享，通过交换和买卖以生成大数据，在此之上探索全新的产品和服务。

13.1.2 大数据时代下金融业的机遇和面临的挑战

金融业是最重视信息科技的行业之一，但是大数据时代猝然来临也让金融业措手不及。大型的电子商务公司在小额支付、小额贷款、供应链金融等领域突飞猛进的发展，甚至让大型银行都有了切肤之痛。

大数据时代的来临，意味着机遇，也意味着挑战。尽管我们无法准确预判大数据最终会对金融业产生什么影响，但深入研究大数据时代金融业的机遇和挑战，有利于金融行业在大数据时代趋利避害。

1. 大数据时代下金融业的机遇

在大数据时代，金融行业主要有 4 方面的机遇，如表 13-1 所示。

表 13-1 金融行业在大数据时代的机遇

机　　遇	说　　明
拓宽行业发展空间	满足客户需求是金融企业生存和发展的前提，大数据和互联网的发展使金融业能够更好地满足客户需求。大数据技术在营销领域的应用将能更有效地发现客户和客户的潜在需求，进行精准营销，特别是投资理财中标准化产品的营销。大数据和互联网的运用也有利于改善消费者的用户体验，提高消费者满意度，改善行业形象
提高行业风险管理能力	大数据技术在风险管理领域的应用将支持金融业更精准的定价原则，提高投资风险识别能力，提升金融业的风险管理能力和水平。以精算为例，大数据有利于扩大用于估算风险概率的数据样本，从而提升精算的准确度，有利于收集更加多维全面的数据，从而形成更加科学的精算模型，也有利于把整体数据样本进一步细分为子样本，为精准定价提供精算基础
提升行业差异化竞争能力	大数据通过对客户消费行为模式的分析，提高客户转化率，开发出不同的理财产品，满足不同客户的市场需求，实现差异化竞争
提升金融业资金运用水平	大数据基于精确量化的投资分布，可以提升金融机构资产负债管理水平，可以在资本市场实施更精准的风险投资组合策略，提高金融业在资本市场的投资回报水平

2. 大数据时代下金融业面临的挑战

在看到机遇的同时，必须看到大数据时代金融业还面临一些严峻挑战，如表 13-2 所示。

表 13-2 金融行业在大数据时代的挑战

挑战	说明
思维方式面临冲击	虽然我国金融市场不断涌现创新产品，但总体上是延续了发达金融市场发展的脉络。但大数据对思维方式的冲击可能是颠覆性的。例如，“阿里小贷”对银行的影响给我们很多启示。在技术剧烈变化的条件下，如果思维方式跟不上，企业经营或资金监管都可能出大问题
数据基础比较薄弱	这些年，金融业在大数据战略和网络经营等方面进行了积极探索，但总体上保险业大数据的基础还很弱，和互联网等行业相比差距很大。同时，不同主体间大数据应用能力存在较大差异。各大金融主体挖掘内部数据，收集外部数据，对数据分析和处理，发现数据背后价值的能力良莠不齐，这将直接影响金融市场核心竞争力
外部竞争可能加剧	在大数据时代，与拥有数据的信息产业相比，金融业将处于相对不利的市场地位，金融业面临来自互联网企业和科技公司业务分割的竞争压力，金融行业的生存空间受到挤压，其竞争力可能弱化
人才储备严重不足	现在，高端信息技术人才匮乏是制约金融业发展的重要因素之一，在大数据时代，金融业在人才上的问题显得更加突出

13.1.3 金融业该如何“迎战”大数据

IT 技术和金融产业，貌似是两个完全不相同的领域，却隐藏着密切的联系。大数据处理作为时下最热门的 IT 技术之一，随着数据仓库、数据安全、数据分析以及数据挖掘等等围绕大数据的商业价值的利用逐渐成为业内人士争相谈论的利润焦点。在这些纷繁杂乱的数据背后，它能找到更符合用户兴趣爱好的产品与服务，并实时对产品与服务进行跟踪性的调整和优化，这就是大数据对我们所带来的影响，从而更进一步地影响着各个行业。

因此，大数据必然引发金融行业的重要变革，金融业应在战略层面重视大数据时代的到来，并以此为契机提升金融行业的创新能力、服务能力和风险管理能力，完善保险监管体系，如表 13-3 所示。

表 13-3 金融业在大数据时代的战略

发展战略	具体说明
建立适应大数据时代要求的数据治理架构	金融企业要结合自身的实际需求，研究制定大数据战略，统筹规划大数据应用，主要表现在以下 3 个方面： ➢ 营造数据文化。将现有数据转化为信息资源，让决策更加有的放矢，让发展更加贴近市场

续表

发 展 战 略	具 体 说 明
建立适应大数据时代要求的数据治理架构	➢ 有效管理数据。进一步健全数据管理决策机制和内部协调机制，提高数据管理制度的可操作性和执行力 ➢ 挖掘监管数据。要提高数据采集能力、分析能力和使用能力，把大量沉睡的数据变为有利于改进监管的信息，为实施动态监管、过程监管和实时监管，提升监管的针对性和有效性提供数据和技术支撑
利用大数据技术开发更多金融产品	大数据处理技术的运用，可以给金融企业提供全新的、更多的业务品种。大数据处理技术的运用，可以帮助金融机构根据客户的习惯、喜好，开发更多适合客户的个性化产品，实现“一对一”的自助服务
加快建设适应大数据时代要求的信息化基础	实现大数据运用的根本和前提是基础设施建设。在大数据时代，必然要求金融机构增加信息化基础设施投入，这样才更易于数据的整合与集中、扩展与伸缩、管理与维护，同时基础设施要具备极高的可靠性、可控性和安全性。为此，金融业必须要建立适应大数据时代要求的信息化基础架构，搭建基础数据技术平台。要统筹好历史数据和当前采集数据的关系，统筹好大数据背景下精算技术、统计技术和数据挖掘技术的融合，统筹好结构化数据和非结构化数据的采集、分析和使用，充分挖掘历史积累保险数据的潜在价值，积极学习运用大数据技术提升分析现实数据的能力
利用大数据技术改善银行客户关系	要有针对性地改进客户服务，就必须了解客户的潜在需求，对客户的维护过程进行及时的响应。金融行业对数据的存储要求特别高，诸如银行、证券、保险等金融领域，每天都会产生大量的数据，这些数据都会被一一存放在交易系统里，金融机构要做的努力就是对这些数据进行深入的挖掘和全面的分析，从而大大提高工作效率和风险防范能力，进而改进客户服务，提升金融行业的盈利水平。例如，银行可以通过结构化数据为客户提供服务，根据客户的交易信息、历史记录来分析客户的理财习惯。通过借助大数据处理技术，使金融行业的服务具有“3A”特性，即 Anytime（任何时候）、Anywhere（任何地方）、Anyhow（以任何方式）为客户提供金融服务，从而吸引和留住更多的优质客户，扩大客户群，开辟新的盈利增长点
进一步加强与互联网公司、数据公司的合作	互联网公司和数据公司既是金融业发展的重要参与者，也是金融市场主体合作共赢的重要对象。大数据时代对金融业驾驭数据能力提出了更高的要求。金融市场主体不仅要收集行业的内部数据，更要依靠互联网公司和数据公司收集外部数据。金融机构要切实加强同互联网公司和数据公司的战略合作，提高内外部数据信息的整合能力
防范大数据时代的信息安全风险	大数据意味着来自多方面的海量数据，也意味着数据处理软硬件环境更加复杂。集中的数据更复杂、更敏感，更易成为攻击者的目标，常规的安全管理策略，已无法满足安全要求。各金融机构都要严格遵守监管机构和信息化主管部门制定的规章制度，进一步完善信息化治理，强化责任落实，加强信息安全培训，提升信息安全意识，完善信息安全预警和应急响应机制，进一步健全与大数据时代相适应的信息安全保障体系

续表

发展战略	具体说明
引进与培养金融业大数据专业人才	数据科学是一门交叉学科，涉及数学、统计学、计算机科学、数据可视化技术以及各领域专业知识。大数据的运用，关键还是人才。无论是基础建设，还是数据分析与系统维护，都需要专业的数据人才。各金融机构要加大力气，舍得投入，抓好大数据人才的引进与培养，打造一支数量充足、结构合理、素质优良、表现卓越的复合型专业人才团队
创造良好的大数据时代监管环境	大数据时代给金融行业发展带来深刻影响的同时，也对金融监管制度提出更高的要求。金融监管机构要顺应大数据时代的潮流，为行业创新发展营造良好环境，主要从以下 4 个方面做起： ➢ 强化基础建设。建立大数据质量标准，消除壁垒，推进信息共享，建立信息隐私保护制度，加强信息安全的保护，建立安全有效的大数据共享使用环境 ➢ 鼓励包容创新。以开放的心态，支持金融机构运用大数据进行产品、服务、管理等方面的有益创新，并在监管上及时跟进 ➢ 完善监管制度。对金融市场基于大数据的新事物新探索，适时制定监管制度加以规范，减少监管死角和监管真空，保护消费者合法权益，同时也要避免过度监管 ➢ 注意创新风险。加强对风险的预警跟踪，对大数据条件下的新风险保持足够的敏感和警惕，促进金融市场可持续健康发展
有效利用大数据技术提升金融机构的服务效率	金融行业要想不断发展就离不开大数据处理技术，大数据处理技术在存储和处理结构框架等方面的优势将帮助金融行业充分掌握业务数据的价值，降低运营成本，发掘新的盈利模式，为客户提供更为全面、贴心的金融服务。金融行业必须始终坚持“以客户为中心”的服务理念，以“大数据处理技术”作为支撑，满足客户的多样化需求，实现客户服务的最大价值

笔者认为，使用大数据金融的概念，制定并实施大数据金融战略，更能体现金融业自身的实力和潜力，也更能与网络业及其他行业有机融合，平等竞争，在大数据时代找到自身生存发展的机会也更大。

13.2 金融行业大数据应用案例

如今，金融业面临众多前所未有的跨界竞争对手，市场格局、业务流程将发生巨大改变，企业更替兴衰；未来的金融业，业务就是 IT，IT 就是业务；金融业将开展新一轮围绕大数据、移动化、云的 IT 建设投资。本节主要介绍金融行业大数据的应用案例，希望对读者有一定的启发和学习价值。

13.2.1 【案例】淘宝网掘金大数据金融市场

随着国内网购市场的迅速发展，淘宝网等众多网购网站的市场争夺战也进入白热化状态，网络购物网站也开始推出越来越多的特色产品和服务。

1. 余额宝

以余额宝为代表的互联网金融产品在 2013 年刮起一股旋风，截至目前，规模超 1000 亿元，用户近 3000 万，如图 13-1 所示。相比普通的货币基金，余额宝鲜明的特色当属大数据。以基金的申购、赎回预测为例，基于淘宝和支付宝的数据平台，可以及时把握申购、赎回变动信息。另外，利用历史数据的积累可把握客户的行为规律。





图 13-1 余额宝手机端界面

2. 淘宝信用贷款

淘宝网在聚划算平台推出了一个奇怪的团购“商品”——淘宝信用贷款。开团不到 10 分钟，500 位淘宝卖家就让这一团购“爆团”。他们有望分享总额约 3000 万元的淘宝信用贷款，并能享受贷款利息 7.5 折的优惠。据悉，目前已经有近两万名淘宝卖家申请过淘宝信用贷款，贷款总额超过 14 亿元。

淘宝信用贷款是阿里金融旗下专门针对淘宝卖家进行金融支持的贷款产品。淘宝平台通过以卖家在淘宝网上的网络行为数据做一个综合的授信评分，卖家纯凭信用拿贷款，无需抵押物，无需担保人。由于其非常吻合中小卖家的资金需求，且重视信用无担保、抵押的门槛，更加上其申请流程非常便捷，仅需要线上申请，几分钟内就能获贷，被不少卖家戏称为“史上最轻松的贷款”，也成为淘宝网上众多卖家进行资金周转的重

要手段。

3. 阿里小贷

淘宝网的“阿里小贷”更是得益于大数据，它依托阿里巴巴（B2B）、淘宝、支付宝等平台数据，不仅可有效识别和分散风险，提供更有针对性、多样化的服务，而且批量化、流水化的作业使得交易成本大幅下降。

每天，海量的交易和数据在阿里的平台上跑着，阿里通过对商户最近 100 天的数据分析，就能知道哪些商户可能存在资金问题，此时的阿里贷款平台就有可能出马，同潜在的贷款对象进行沟通。

【案例解析】：通常来说，数据比文字更真实，更能反映一个公司的正常运营情况。通过海量的分析得出企业的经营情况，这就是大数据的应用。在本案例中，正像淘宝信用贷款所体现的那样，这种新型微贷技术不依赖抵押、担保，而是看重企业的信用，同时通过数据的运算来评核企业的信用，这不仅降低了申请贷款的门槛，也极大简化了申请贷款的流程，使其有了完全在互联网上作业的可能性。

大数据的价值已经得到互联网公司以及金融机构的认可，笔者认为：“谁掌握的‘拼图’图块多，谁就能快速拼出客户的图谱，成为真正的王者。”然而，目前来看，谁都不愿意轻易地交出自己手上的“拼图”，于是，互联网公司、银行、支付机构等各个海量数据的拥有者展开了激烈的金融数据争夺战。

13.2.2 【案例】IBM 用大数据预测股价走势

不久前，IBM 使用大数据信息技术成功开发了“经济指标预测系统”。借助该预测系统，可通过统计分析新闻中出现的单词等信息来预测股价等走势。

IBM 的“经济指标预测系统”首先从互联网上的新闻中搜索与“新订单”等与经济指标有关的单词，然后结合其他相关经济数据的历史数据分析与股价的关系，从而得出预测结果。

在“经济指标预测系统”的开发过程中，IBM 还进行了一系列的验证工作。IBM 以美国“ISM 制造业采购经理人指数”为对象进行了验证试验，该指数以制造业中的大约 20 个行业、300 多家公司的采购负责人为对象，调查新订单和雇员等情况之后计算得出。实验前，首先假设“受访者受到了新闻报道的影响”，然后分别计算出约 30 万条财经类新闻中出现的“新订单”、“生产”以及“雇员”等 5 个关键词的数量。追踪这些关键词在这段时期内的搜索数据变化情况，并将数据和道指的走势进行对比，从而预测该指数的未来动态。

IBM 研究称，一般而言，当“股票”、“营收”等金融词汇的搜索量下降时，道指随后将上涨，而当这些金融词汇的搜索量上升时，道指在随后的几周内将下跌。

据悉，IBM 的试验仅用了 6 小时，就计算出了分析师需要花费数日才能得出的预测值，而且预测精度几乎一样。

【案例解析】：从本案例可以看出，大数据不再仅仅局限在媒体与厂商之间的讨论，它犹如一场数据旋风开始席卷全球，从各行各业的 IT 主管到政府部门都开始重视大数据及其价值。

目前，不少信息系统企业都在使用大数据信息技术开发预测系统。例如，2011 年，英国对冲基金 Derwent Capital Markets 建立了规模为 4000 万美金的对冲基金，该基金是首家基于社交网络的对冲基金，该基金通过从 Twitter 的数据内容来感知市场情绪，从而进行投资。无独有偶，美国加州大学河滨分校也公布了一项通过对 Twitter 消息进行分析从而预测股票涨跌的研究报告。

笔者认为："企业数据就是新时代还未开采的石油，具有非常之高的价值。"国外一些金融机构已经开始做一些前瞻性的研究了，这种做法是非常值得国内金融机构学习和借鉴的。例如，国内大部分证券公司仍然没有摆脱交易性数据为主的特点，但很多有前瞻意识的证券公司已经开始做一些转型了，对微博、互联网等外部数据进行一些分析与预测。

13.2.3 【案例】汇丰银行采用 SAS 管理风险

近日，汇丰银行选择 SAS 防欺诈管理解决方案构建其全球业务网络的防欺诈管理系统。据悉，这一解决方案是一种实时欺诈防范侦测系统。

SAS 被誉为"全球 500 强背后的管理大师"，是全球领先的商业分析软件与服务供应商。SAS 通过三部分服务（包括软件及解决方案服务、咨询服务、培训及技术支持服务）帮助客户洞察商机，成就变革，改善业绩。

凭借丰富的行业专业知识，SAS 的行业解决方案在各领域为行业解析蕴藏于信息之中的独特的商业问题。例如金融服务领域的信用风险管理问题、生命科学领域加快药物上市速度和识别零售领域的交叉销售机会等问题。SAS 还提供跨职能解决方案，不分行业地帮助企业克服其面临的挑战。例如增加客户关系价值、测量和管理风险、检测欺诈和优化 IT 网络等。

汇丰银行与 SAS 在防范信用卡和借记卡欺诈的基础上，共同扩展了 SAS 防欺诈管理解决方案的功能，为多种业务线和渠道提供完善的欺诈防范系统。这些增强功能有助于全面监控客户、账户和渠道业务活动，进一步提高分行交易、银行转账和在线付款欺诈以及内部欺诈的防范能力。通过监控客户行为，汇丰银行可以优化并更加有效地利用侦测资源。

汇丰银行利用 SAS 系统，通过收集和分析大数据解决复杂问题，并获得非常精确的洞察，以加快信息获取速度和超越竞争对手。因此，汇丰银行还将继续采用 SAS 告警管

理、例程和队列优先级软件，提高运营效率，以便迅速启动紧急告警。

【案例解析】：在当今这个海量数据的时代，如何找到大数据中蕴含的前所未有的商业价值？笔者认为高性能分析就是那把“钥匙”。在本案例中，SAS 高性能分析可以帮助用户：将相关的大数据转变为真正的商业价值，采用世界顶级的分析技术来生成精确的洞察，快速获得答案来改变企业的运营模式，以及部署一个适合未来扩展的分析架构。

总之，高性能分析环境让用户可以充分利用 IT 投资，同时克服原有架构的约束，从大数据资产中产生高价值的洞察。

13.2.4 【案例】Kabbage 用大数据开辟新路径

Kabbage 是一家为网店店主提供营运资金贷款服务的创业公司，总部位于美国亚特兰大，截至目前已经成功融资六千多万美元。Kabbage 的主要目标客户是 eBay、亚马逊、雅虎、Etsy、Shopify、Magento、PayPal 上的美国网商。

Kabbage 与“阿里小贷”的经营模式类似，通过查看网店店主的销售和信用记录、顾客流量、评论以及商品价格和存货等信息，来最终确定是否为他们提供贷款以及贷多少金额，贷款金额上限为 4 万美元。店主可以主动在自己的 Kabbage 账户中添加新的信息，以增加获得贷款的概率。Kabbage 通过支付工具 PayPal 的支付 API 来为网店店主提供资金贷款，这种贷款资金到账的速度相当快，最快十分钟就可以搞定。

Kabbage 用于贷款判断的支撑数据的来源除了网上搜索和查看外，还来自于网上商家的自主提供，且提供的数据多少直接影响着最终的贷款情况。同时，Kabbage 也通过与物流公司 UPS、财务管理软件公司 Intuit 合作，扩充数据来源渠道。

目前，使用 Kabbage 贷款服务的网店店主已达近万家，Kabbage 的服务范围目前仅限于美国境内，不过公司打算利用这轮融资将服务拓展至其他国家。

【案例解析】：基于大数据的商业模式创新过程有两个核心环节：一是数据获取；二是数据的分析利用。在本案例中，Kabbage 与阿里金融的区别在于数据获取方面，前者是从多元化的渠道收集数据，后者则是借助旗下平台的数据积累，其中网上商家可自主提供数据且其数据的多少直接决定着最终的贷款额度与成本，这充分体现出大数据的资产价值，就如同传统的抵押物一样可以换取资金。

笔者觉得，虽说大数据是一座极具价值的“金矿”，但如果不能科学地加以利用，那么大数据就变成了一堆堆毫无用处的“石头”，Kabbage 就是借助大数据技术，并结合金融行业的特点，有效地控制了风险，实现了完美融合和创新。

金融是服务于实体经济的，随着大数据时代的到来，传统的实体经济形态正在向融合经济形态转变，同时虚拟经济也快速兴起，金融的服务对象必将随之发生变化，这种转变为金融业带来了巨大的机遇和挑战，如图 13-2 所示。



图 13-2　融合经济产生新的金融需求

专 家 提 醒

虚拟经济（Fictitious Economy）是经济虚拟化（西方称之为“金融深化”）的必然产物，是指基于计算机和互联网产生的一种经济形态，其产品和服务都具有虚拟化的特点，具体包括软件、网络游戏、社交网络、搜索引擎、门户网站等细分市场领域。实体经济是指物质的、精神的产品和服务的生产、流通等经济活动。随着新兴信息技术的快速发展，实体经济与虚拟经济正在加速融合，从而衍生了未来的主体经济形态，即融合经济，电子商务、O2O 模式都是融合经济发展进程的一个产物。

13.2.5　【案例】大数据时代信用卡该怎么玩

中信银行信用卡中心是国内银行业为数不多的几家分行级信用卡专营机构之一，也是国内最具竞争力的股份制商业银行信用卡中心之一。近年来，中信银行信用卡中心的发卡量迅速增长。

2013 年 11 月，在中信银行与腾讯联合发布“中信银行 QQ 彩贝联名信用卡”仪式上，中信银行信用卡中心总裁陈劲表示，该行信用卡发卡量已突破 2000 万张，未来将充分利用互联网基因和大数据技术挖掘客户需求。

过去，中信银行信用卡中心无论在数据存储、系统维护等方面，还是在有效地利用客户数据方面，都面临巨大的压力。同时，为了应对激烈的市场竞争，中信银行信用卡中心迫切需要一个可扩展、高性能的数据仓库解决方案，支持其数据分析战略，提升业务的敏捷性。

2010 年 4 月，中信银行信用卡中心实施了 EMC Greenplum 数据仓库解决方案。Greenplum 数据仓库解决方案为中信银行信用卡中心提供了统一的客户视图，借助客户

统一视图，中信银行信用卡中心可以更清楚地了解其客户价值体系，从而能够为客户提供更有针对性和相关性的营销活动。

基于数据仓库，中信银行信用卡中心现在可以从交易、服务、风险、权益等多个层面分析数据。通过提供全面的客户数据，营销团队可以对客户按照低、中、高价值来进行分类，根据银行整体经营策略积极地提供相应的个性化服务。

基于 Greenplum 解决方案在系统维护方面的便捷简单，中信银行信用卡中心每年减少了大约 500 万元的数据库维护成本，这有助于减少解决方案的总拥有成本。

【案例解析】: 在本案例中，Greenplum 解决方案采用了“无共享”的开放平台 MPP 架构，此架构是为 BI 和海量数据分析处理而设计，相比普通的数据库系统，该系统提供了更高的可扩展性。与其他产品相比，Greenplum 解决方案可以给中信银行信用卡中心提供最高级别的性能。同时，该解决方案与银行所使用的硬件、应用程序和数据源实现了有效集成。此外，Greenplum 解决方案通过把数据集中在一个统一的平台，极大地减少了系统维护的工作量。

笔者认为，大数据对信用卡产品的营销具有很大的促进作用。例如，在大数据的环境下，银行可以利用先进的互联网、云计算等新兴技术，对消费者的刷卡行为进行数据化的分类、统计，通过整理数据获取消费者的消费习惯、消费能力、消费偏好等非常重要的数据信息。通过客户数据、财务数据来区隔客户，通过消费区域定位、内容定向，知晓他们的消费习惯，然后进行深入地数据分析挖掘和展开精准营销。

14

交通：畅通无阻的数据

学前提示

坐在家里，打开手机就能知道高架是否拥堵；开车上路，提前几个路口就能收到关于路况的短信提醒……这一切，已经变成现实。大数据的分析和应用还将在道路交通中发挥更大作用。当交通遇上大数据，智能交通便应运而生。

要点展示

- 交通行业大数据解决方案
- 交通行业大数据应用案例

14.1 交通行业大数据解决方案

出门堵车，出租车打不到……每每出门这些烦恼都会困扰着我们，智能交通已经不仅仅是一种畅想，而是每个人都亟待享受到的便利。车驶在路上，人走在街边，不知不觉中他们都成为智能交通中的大数据，“解铃还须系铃人”，智能交通需要大数据来给出答案。

14.1.1 5大日益突出的城市交通难题

随着我国城市人口的增多和汽车的增加，城市交通问题日益突出。在许多大城市，由于过量的汽车，经常导致交通阻塞，交通事故频发，大气遭到污染等。交通问题已经给城市社会经济发展带来了严重影响。如表14-1所示为大城市主要存在的交通问题。

表14-1 大城市主要存在的交通问题

交通问题	产生原因和危害
交通阻塞	人们经常把容易塞车的道路，称为交通瓶颈（或交通樽颈）。相对于道路网的承载力来说，汽车数量过多，是诱发交通阻塞的主要原因。从某种程度上说，交通阻塞是汽车社会的产物。在人们上下班的高峰期，交通阻塞现象尤为明显，在很多大城市中心区，高峰期交通速度每小时仅有16km。交通阻塞导致时间和能源的严重浪费，影响城市经济的发展。在大城市，汽车数量的增长速度远远高于道路的建设速度，道路的建设和汽车的增加有可能形成恶性循环，导致更为严重的交通阻塞
交通事故	交通事故是许多大城市日趋严重的问题。交通事故不但导致了对贵重医疗设施需求的增加，而且使受伤者痛苦不堪。据统计，仅1978年，美国就有52653人死于机动车事故
公共交通	公共交通问题主要表现在以下两个方面： ➢ 由于对公共交通投资不足，致使峰值期人们对公共交通的需求大于供给，造成交通拥挤 ➢ 由于对公共交通的需求波动大，高峰期过于拥挤，而非高峰期使用又不充分，造成收入锐减 由此可见，高峰时间和非高峰时间的公共交通是一对难以解决的矛盾。如果增加投资来满足高峰期人们对公共交通的需求，那么在非高峰时间，这些公共交通设施大部分将处于闲置状态，造成浪费。在发达国家，这种情况一方面对于公共交通工具依赖性较大的低收入阶层是一个打击，另一方面又促进了中产阶级甚至低收入阶层对小汽车的依赖性。这又使公共交通进一步萎缩，形成恶性循环。在发展中国家，则使公共交通高峰时间的拥挤现象更为严重，从而加剧了城市交通问题

续表

交 通 问 题	产生原因和危害
步行者问题（包括非机动车）	步行或骑自行车在目前仍然是一种重要的交通方式，交通量很大。据调查，在伦敦南部，人们上下班之外的行程中，50%以上的人是靠步行。现在，很多城市都在为改善道路交通进行规划，如加宽机动车道，但却很少考虑步行者的需求。例如，在一些城市，为了照顾汽车，人行道变窄了，交通安全岛取消了，不设置横越马路的绿灯信号，机动车辆被允许停放在人行道上或道旁，这些都给步行者带来麻烦和危险。最主要的是，步行者还必须忍受噪声、烟雾、汽油味等污染，严重影响身体健康。现在，很多大城市已开始着手解决步行者问题，如规定在中心商业区一些重要街道上禁止车辆通行，设为步行街或步行区；在市中心除公共汽车外，其他车辆白天均不得通过等，但解决的力度还远远不够
停车困难	当汽车处于静止状态时，就要占据一定空间，汽车越多占据的空间就越大。在城市中心区，人多车多空间少，停车场与汽车数量很不相称，停车也最困难。尽管近十多年来在市区建了许多多层停车场，但仍满足不了停车需求。于是很多城市通过颁布法令，限制在市中心区停车，以控制进入市中心区汽车的数量，但这些措施并没有解决停车问题。因此，如何有效地解决停车问题仍在探讨中

专 家 提 醒

例如，美国政府曾在20世纪70年代中期制定过一个方案，迫使个人使用公共汽车来代替小汽车。但很多人反对这个方案，认为这样会减少家庭小汽车的数量，从而改变消费模式，减少就业机会，会导致失业、福利、职业培训和贫困等问题出现。另外，发展公共交通还需要政府大量补贴，其结果将限制解决其他问题资金的流动，或者被迫增加税率。高税率将使货币从个人手中分配到政府手里，从而可能造成社会经济体系变化，也增加了政治不稳定性。

由此可见，交通问题的解决绝不是一朝一夕的事情。为此，及时、高效、准确获取交通数据是分析交通管理机制，构建合理城市交通管理体系的前提，而这一难题可以通过大数据管理得到解决。

14.1.2 大数据为交通难题开出的药方

大数据时代的到来，为解决交通问题开出了有效“药方”。与传统的数据收集方式不同，云时代的大数据通过对数据实时收集和分析，得以实现个人出行的个性化、方便化和智能化。另外，大数据将海量数据聚合在一起，将离散的数据需求聚合起来形成数据长尾，从而满足传统中难以满足的需求，例如交通需求。

因为无论是交通基础设施、交通运行状态还是交通服务对象和交通运载工具，每时

每刻都在产生着大量的数据，以大数据的思路和角度来看，这些都是正待挖掘的宝藏，能为交通决策和服务带来新的解题思路。面对大数据的浪潮，交通运输行业不应是一个“路人”，而是要敞开胸怀，积极地拥抱和融合，借着大数据的力量高度进行内视和审度，再回首，相信会豁然开朗，柳暗花明。

用大数据管理交通是交通管理模式的变革，与此同时也变革了公共交通市场管理的整个内涵，而阻碍传统交通的瓶颈也可通过大数据解决，如表 14-2 所示。

表 14-2　大数据为交通难题开出的“药方”

具体药方	交通症状	对症下药
大数据可以跨越行政区域的限制	行政区域的划分在促进各个行政区域自治的同时，也导致各个地方政府追求各自辖区利益的最大化，而对地方政府之间边界区的公共交通基础设施、过境交通线路等缺少建设	利用交通大数据的虚拟性，有利于共信息跨越区域管理，只要多方共同遵照相关的信息共享原则，就能在已有的行政区域下解决跨域管理问题
大数据具有信息集成优势和组合效率	大部分城市的各类交通运输管理主体分散在不同主管部门，呈现出条块分割的现象。这种分散造成公共交通管理的碎片化，如交通信息分散、信息内容单一等问题	大数据有助于建立综合性立体的交通信息体系，将用户可能利用的各种交通数据纳入系统，构建公共交通信息集成利用模式，发挥整体性交通功能，通过在大数据中进行集成检索、利用和分析来提取相关信息，满足各种交通需求，以解决实时交通障碍
大数据能较好配置公共交通信息资源	传统的交通部门权责界定未厘清，专业分工的细化也促使公共交通管理部门职能重叠，因而在运营上浪费大量人力、物力	大数据能辅助人们制定出较好的统筹与协调解决方案，在各个交通部门之间合理配置交通职能，针对有关道路问题进行合理信息资源配置
大数据能促进公共交通均衡性发展	用传统的思维来改善交通拥堵，一般是加大基础设施投入，即加宽道路、增加道路里程来提高交通通行能力，但这种做法又会受到土地资源的限制，而且这种解决模式不利于交通发展、城市空间发展以及土地利用发展这三者之间的整合	大数据解决方案可以将技术决定论与制度理论相结合，将信息技术应用于公共交通，从制度层面提高信息资本利用率，减少对诸如土地等外部资源的依赖

目前，世界各地政府也都纷纷将交通运输数据由纸质型转向数字方式储存，建立智能交通系统，人们可查看交通流量计数，也可依据车辆行程和路况拥挤程度进行电子收费，从而对交通堵塞和交通污染排放进行隐形控制。

14.1.3 大数据解决交通难题 4 大优势

及时、高效、准确的交通数据获取是分析交通管理机制，构建合理城市交通管理体系的前提，而这一难题可以通过大数据管理得到解决。总的来说，用大数据解决交通难题具有 4 大优势，如图 14-1 所示。

图 14-1 大数据解决交通难题的 4 大优势

专 家 提 醒

例如，美国仅次于房屋的第二大消费成本就是交通运输，美国司机一年只有 4%的时间在开车，但却要每年为车辆支付 8000 美元。如在新泽西州引入大数据处理交通堵塞问题之前，其主要依赖交通摄像机和耗资 2 万美元的路边传感器，但这些信息仅覆盖整个州道路的 5%。引入 INRIX 大数据管理系统之后，尽管新泽西州每年耗费在 INRIX 系统上的资金要达 45 万美元，但其覆盖面更广，信息准确性更高，而且给人们减少的时间成本都是无法计量的。

14.1.4 如何应用大数据解决交通问题

转型中的交通也面临着调整发展结构、提升发展质量的难题，此时与大数据时代相遇实为幸事，因为大数据为交通难题带来了解决方案。在交通问题解决过程中，基于大

数据的智能交通数据处理体系流程如图 14-2 所示。



图 14-2　基于大数据的智能交通数据处理体系流程

专 家 提 醒

公共交通的智能化管理表现在：一旦某个路段发生问题，能立刻从大数据中调出有用信息，确保交通的连贯性和持续性；另一方面，大数据具有较高预测能力，可降低误报和漏报的概率，可随时针对公共交通的动态情况给予实时监控。

应用大数据解决交通问题的具体流程说明如表 14-3 所示。

表 14-3　应用大数据解决交通问题的具体流程说明

解 决 流 程	具 体 内 容
收集和输入数据	这些数据包括静态数据和动态数据，前者指道路环境、车辆信息等长时间不会改变的数据，这类数据通过线圈（类似于磁性检测器）和摄像机（交通视频）进行搜集；后者指在交通运行中而产生的实时数据（如车辆行驶速度），这类数据通过 GPS 全球定位技术、手机网络信号来搜集
交换和处理数据	数据中心对实时交通流数据进行提取，同时规定统一的数据格式，从而促进数据交换中心之间对数据进行交换和处理
储存和集成数据	通过基于云计算的云存储来对数据进行储存，将大数据集成起来
管理和运用数据	控制中心将这些大数据在电脑地图上以不同色彩来呈现，分别以不同颜色注明各个路段的拥堵情况，如图 14-3 所示。在这一体系中，为了真正利用好大数据，必须要处理好如下问题：高速连接、大数据管理、开放数据等

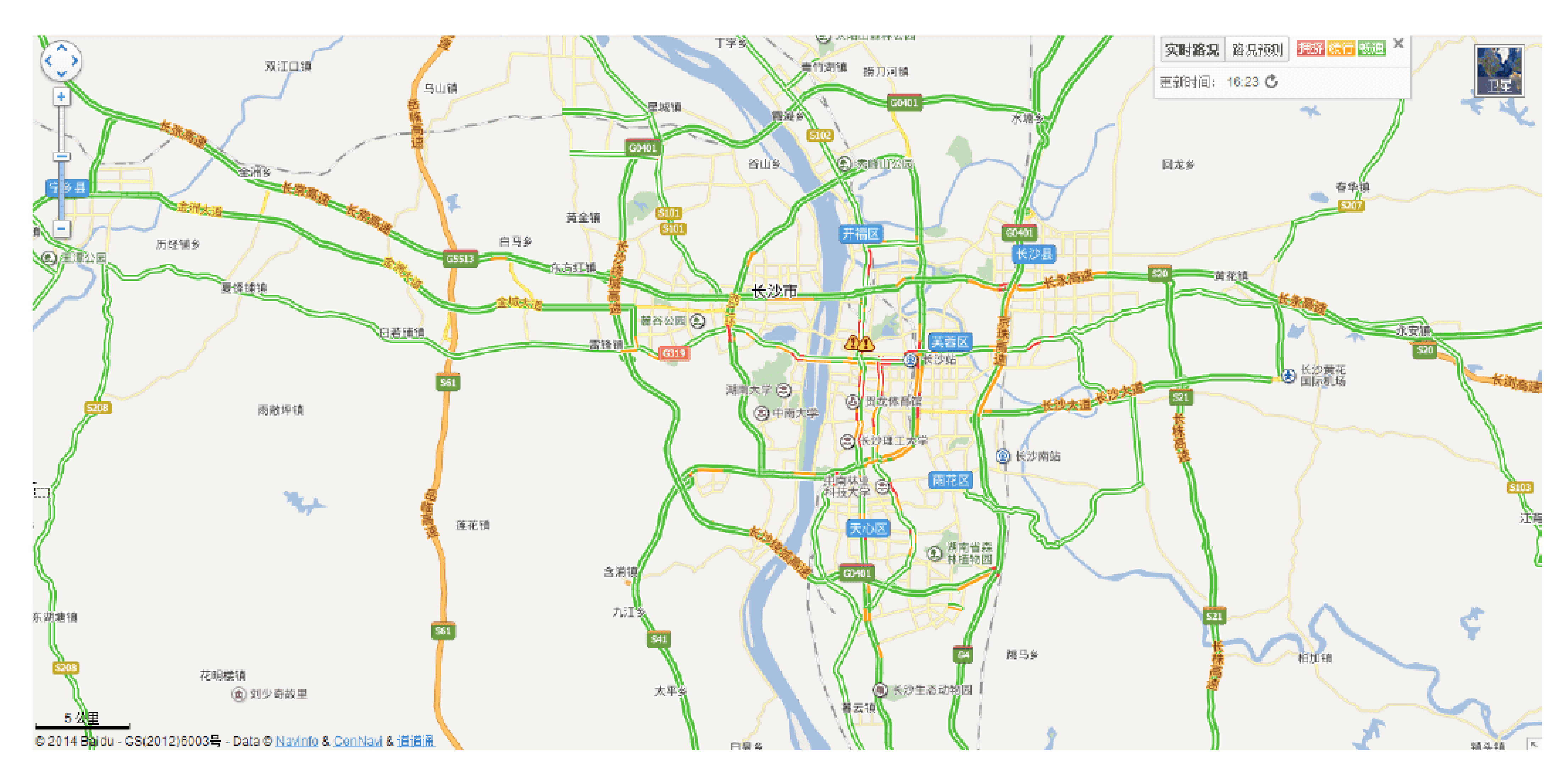


图 14-3　电脑地图上的各个路段的拥堵情况显示

14.1.5　大数据在智能交通行业的挑战

随着信息通信技术的发展，交通运输从数据贫乏的困境转向数据丰富的环境，而面对众多的交通数据，如何从中根据用户需求提取有效数据成为关键所在。大数据管理是一个巨大的挑战，一方面要及时提取交通数据以满足用户需求，另一方面须在数据的潜在价值与个人隐私之间进行平衡。

大数据在智能交通行业面临的挑战与建议如表 14-4 所示。

笔者认为，要真正利用大数据构建一体化的公共交通管理体制，还需要对交通数据采集、处理等方面进行梳理，需要对智能交通系统的构建以及用户界面的完善做进一步研究。总之，数据是智能交通的核心，对交通数据深度处理与分析是其中的关键。

表 14-4　大数据在智能交通行业面临的挑战与建议

面 临 挑 战	具 体 表 现	建 议 方 案
如何开放公共交通数据	目前，大多城市是在私人数据库中管理它们的交通和运输数据，且仅由市政工作人员监视系统性能以及实施改善措施。这种对数据的封闭式管理不会促进信息的增值	交通主管部门应建立诸如 Transportation InformationGroup 的开放交通运输数据的门户网站，尽可能以 XML、Text/CSV、KML/KMZ、Feeds、XLS 等多种格式开放交通运输数据，提高机器可读性；同时，在门户网站上配备数据挖掘和抽取工具，促进用户根据个人喜欢来获取数据；制定促进交通数据共享的奖励措施，推动交通信息的开放和整合

续表

面临挑战	具体表现	建议方案
个人隐私问题	大数据扩大了信息范围，加快了信息传递和共享速度，若不加以严格控制，其所含的商业信息或私密信息就可能泄露，例如个人所在位置、个人出行习惯以及用户最喜欢的主路线等。一旦个人察觉到这些私密信息有泄露，就会抵制大数据管理系统的广泛应用	政府应制定一部完整的数据隐私法，对个人数据的定义、数据可发布范围、数据发布的基本原则、数据可利用的范畴等方面进行规范。交通主管部门在遵守这部法律的基础上，进一步细化可发布的交通信息，并开展数据隐私、安全的教育项目，加大用户对隐私规则的了解。主要原则是：数据的商业性开发、公益性利用能够与个人隐私权之间相平衡，政府在赋予企业更大程度利用数据的权利和获得潜在商业利润的同时，要减少公民对个人隐私和数据安全的担忧
交通数据的存取方式	如今，各地交通机构都具有交通数据并能被大数据管理系统应用，但很多车辆计数（计算交通车辆数目）的数据都以静态格式（如 PDF）存储，使得系统所具备的计数特性无法被除人之外的事物进行检索，这种传统“人对物”的互联网连接方式不符合物联网的“物对物”特性	交通部门必须聚合各种交通数据，一方面要重视数字化交通数据，另一方面要对重要核心交通数据进行纸质保存，这样可以通过资源共享的方式来丰富整个智能交通的数据长尾。此外，为了真正实现公共交通的智能化，可以加大交通数据中心的自动化程度，让用户能自动收发交通数据

14.2　交通行业大数据应用案例

无论在哪里，城市管理者都希望打造畅通、清洁、安全的交通环境，但是凭借印象、推测做出的决策往往经不起实践的检验，一味拓宽道路和盲目规划也会激化人地矛盾。而在大数据时代，数据的分析为交通科学决策和管理提供了一条便捷又较为可行的道路。

14.2.1　【案例】大数据解决波士顿堵车难题

据悉，波士顿可能是美国交通最拥堵的 10 个城市之一，为了解决这个问题，IBM 公司的工程师为波士顿政府建立了一套应用程序，其能将从手机加速器到社交网站上的数据整合在一起，绘制出波士顿交通情况全面而完整的实时图像，供有关人员参考。该方案资金来自 IBM 智慧城市项目，IBM 的 6 位数据分析工程师准备通过整合、分析现有

交通数据，以及来自社交媒体（Twitter）的新数据源，来医治波士顿的交通恶瘤。

在波士顿，每秒钟都有数以百万计的数据点信息，包括 GPS 和手机，这些数据经过分析处理后可以提供交通智能信息。IBM 的专家们以及来自波士顿大学的技术人员准备制定一个优化的交通管理计划，以便更快地发现拥堵问题；通过制定更好的自行车、泊车和交通管理政策，大幅降低碳排放。

IBM 安装在 iPhone 上的移动应用分析软件，类似移动 BI 仪表盘，可供市政规划人员使用，但波士顿市政府透露将来也会发布面向公众的 iPhone 交通应用，将部分数据公开。这些数据包括市政网联网能够实时采集的交通信号灯、二氧化碳传感器甚至汽车的数据，这些数据能够帮助乘客重新调整路线，节省时间和汽油，如图 14-4 所示。

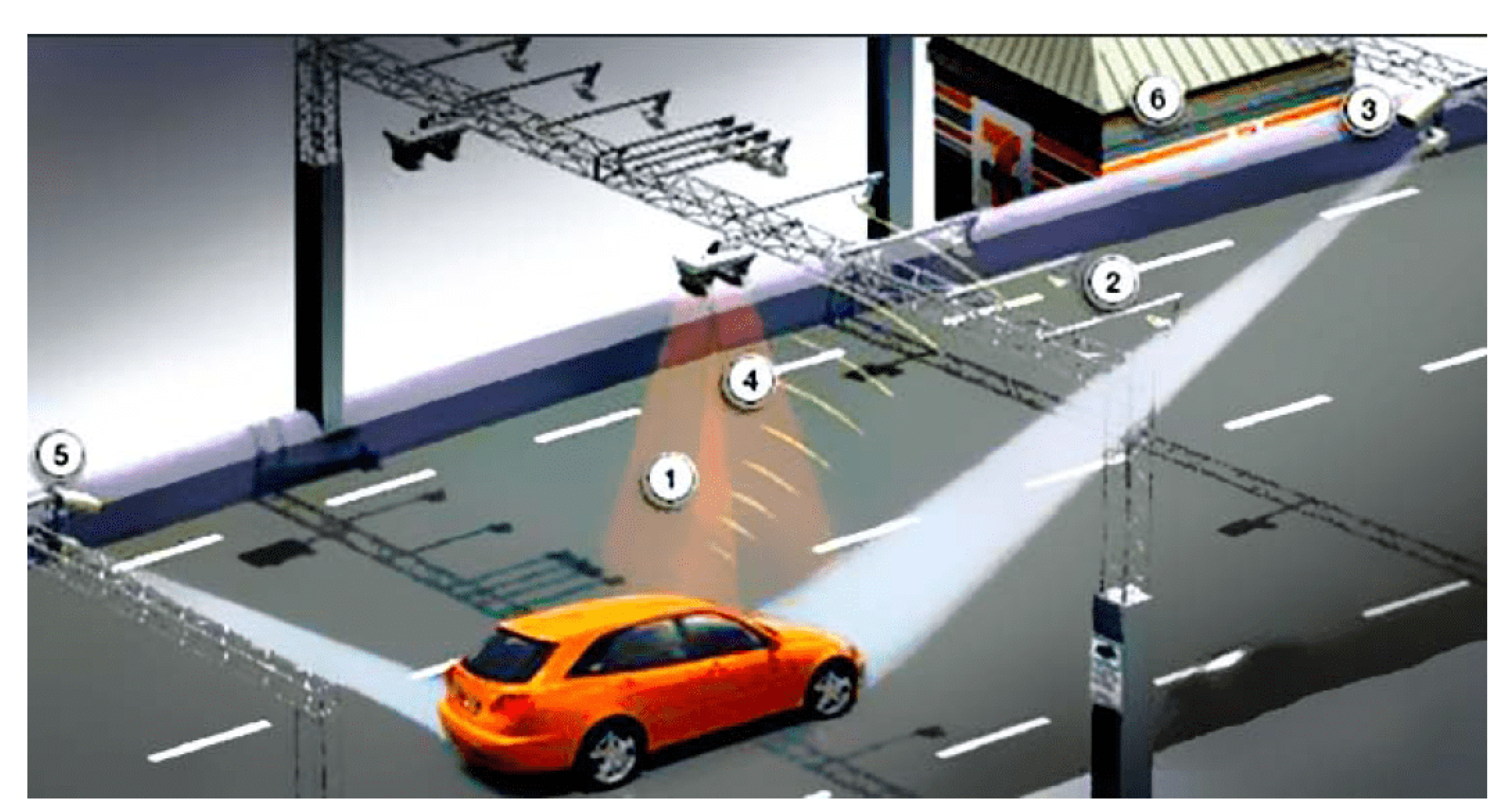


图 14-4　在道路上利用交通信号灯、二氧化碳传感器等采集交通数据

据该预测系统的开发小组——IBM 智能出行者（Smarter Traveler）的项目经理 John Day 介绍，该系统包含三个部分。

第一部分是拥有具有 GPS 功能智能手机的驾驶员用户数据库，该手机可以自动将他们的位置发送到道路网络上，可以让系统掌握驾驶员常常行驶的路线。系统通过查看驾驶员的目的地来判断其常常行走的路线，还会通过道路传感器来收集交通数据。这些传感器包括分布在各大道路上的感应线圈式探测器——一种磁场感应装置，每 30 秒感应一次并汇报车辆通过的信息。

第二部分是 IBM 的交通预测工具（TPT），它是一种通过历史数据来实时预测未来可能发生事件的学习和分析引擎。交通预测工具通过对交通数据的分析来确定较小道路事故与较大交通事故之间的关联。该系统在事故发生的时候会识别出异常情况，然后迅速判断出接下来可能发生的交通模式。

第三部分需要将出行建议发给用户。这时 TPT 已经完成其工作，在用户可能会行走的路线与该路线上可能会存在的问题之间找到了某种关联。与此同时，系统还会通过对交通信号配时、匝道信号控制以及路线规划的改进来帮助用户和交通系统部门在拥堵发生之前可以更好地预测并减少追尾事故的发生。

另外，该程序有望通过跟踪针对同一地点的不同数据流来让有关人员实时调整城市的交通流动情况，或许甚至能调整交通信号灯的模式以避免一些有可能会发生的事故，例如，隧道内的车祸或者球场附近的交通拥堵等。

【案例解析】：针对城市交通堵塞，人们普遍会使用谷歌、微软等技术公司研制的“实时路况”软件了解交通状况，然而很多时候，等到人们发现前方有堵车时，已经为时过晚，他们已经深陷车流中，来不及改道了。在本案例中，如果移动应用分析软件可以对每个城市居民开放，这样大家都可以使用这类整体性的数据分析，更好地制定自己的出行计划。

回到交通的问题，除了不堵车，交通管理对于企业运营和城市构建都有重要意义。例如，企业运输原料，物资在路上耽误的时间越长效率越低，制造的污染和能耗也越多越高。通过对不同行业的交通数据跟踪，政府可以更好地计划和管理企业，有意识地设计产业布局，从而构筑城市可持续核心竞争力。

专家提醒

IBM 公司和加州交通局开发的一个“堵车预警系统”会收集每辆汽车的 GPS 信息，通过数学模型，在堵车尚未发生时便可以预测出哪儿会发生拥堵，市民们甚至可以提前多达 40 分钟便得知交通路况。另外，IBM 公司和加利福尼亚州交通局以及加州大学伯克利分校的创新交通中心合力设计了一款名叫“聪明出行”的系统，它可以让司机们在交通堵塞还没发生之前就预测到哪儿会堵车，它会为用户规划数条出行路线，并用不同颜色呈现它们在可预见的时间内的交通状况。

14.2.2 【案例】谷歌街景带你在家环游世界

谷歌街景（Street View）让科幻小说中的瞬间移动（Teleportation）成为了现实，现在只需轻点鼠标，人们就能实现“远途旅行”。随着全球化和人员流动的加剧，人们希望尽快对一个陌生地区熟悉起来的意愿，为谷歌街景这项新技术提供了广阔的前景。Google 的最终目标是提供全世界的街头景观。

对于不少人来说，能够在世界各地自由穿梭，而不需要真的进行“实体”旅行，这实在算得上是一个伟大的成就。无需经过严酷的穿越，就能够探索数千英里之外的物理空间，听上去就和科幻小说的情节一样梦幻。而现在，谷歌街景已经让人们离瞬间移动的目标更近一步——只是，当然，它不能真的对实体物品进行转移。

谷歌街景是应用于 Google Maps 和 Google Earth 的一项技术，提供世界上许多街道不同位置的全景展现。谷歌街景诞生于 2007 年 5 月 25 日，最初只在美国的几个城市使用，此后逐步扩大到更多的城市和乡村以及更多的国家和地区。

谷歌街景显示的图像是由经过特别改装的车队拍摄的，对于不能行车的地区，如行人专用区、狭窄的街道、小巷和滑雪胜地等，则用三轮车或滑雪车来拍摄。在这些车辆上各有 9 个 360 度全景定向相机，高度约 2.5 米，另外配有全球定位仪和三台激光测距仪用来扫描车头前 180 度范围内 50 米内的物体，还有天线扫描、3G/GSM 和 WiFi 热点等。

“谷歌街景”服务只是谷歌的地图服务的补充，谷歌公司希望用户将它和之前发布的“谷歌地球”结合起来，从而充分了解地球上的每一个地区，如图 14-5 所示。在这些精确定位的地球照片上，不仅仅可以看到哪一户家庭的后院有游泳池或者网球场，以及家门口的汽车型号和颜色，甚至花园里的设施和其中晒日光浴的人也能一览无余。



图 14-5　谷歌街景地图

谷歌地图日前推出了一项全新的街景功能，用户通过一个地图扩展包将可以使用全新“水下街景”功能，畅游谷歌所选取的 6 个海底特定区域的 360 度全景地图。

“水下街景”不仅仅能够为脆弱的且不断在发生变化的海底世界保留珍贵的图片，而且还可以为那些没有机会亲身经历海底世界的用户提供一个身临其境的体验机会。据悉，用户使用“水下街景”功能看到的景象主要是澳大利亚、夏威夷以及菲律宾海域的

珊瑚礁以及生活在其中各种各样的海洋生物。

“Google 街景”自提供服务以来，就一直备受关注与议论。反对者指控 Google 街景地图曝光了太多的个人隐私，有可能侵犯个人隐私，Google 也采取了一系列应对措施，例如对路人脸部做模糊处理，删除一些敏感图片等。

专家提醒

Google 的街景服务采集车在澳大利亚行驶时还顺带收集了道路上的 WiFi 接入点，通过记录网络接入点的信息，可以在没有 GPS 的情况下通过 WiFi 接入点估算用户所在的位置，提供定位服务。但麻烦的是，部分数据被用于其他用途，街景采集车不仅收集 WiFi 接入点数据，并且还记录了 WiFi 网络传送的数据包，如果街景采集车通过一个未经加密的 WiFi 网络，这些数据就会被记录在案，这些数据包中包含电子邮件、用户名与密码等信息。

【案例解析】：在本案例中可以看到，谷歌在收集数据时强调扩展性，毫无疑问其是做得最好的公司之一。谷歌街景采集的数据之所以具有可扩展性，是因为谷歌不仅将其用于基本用途，而且进行了大量的二次使用。GPS 数据不仅优化了其地图服务，而且对谷歌自动驾驶汽车的运作功不可没。

在谷歌街景中，虽然你也许无法做到幻影移形，但你的心可以漫游到你想去的任何地方，此刻的世界就好像真的成了一个地球村。这种连接人与人的方式，是其他任何一种技术无法企及的。

14.2.3 【案例】腾讯 SOSO 让地图更“真实”

腾讯 SOSO 地图于 2011 年 12 月 26 日推出了 SOSO 地图街景服务，这是 SOSO 地图服务增加的一项新功能，其可显示所选城市街道的 360 度全景图像。同时，这也是中国国内第一家提供高清街景地图的公司，受到了媒体、行业及用户的广泛关注。

在制作 SOSO 街景地图时，腾讯并没有自己去采集数据，而是采用订单制，由第三方采集公司来完成。据悉，街景地图一年采集的数据量高达 1PB，光是整理硬盘，搜搜就专门配备了两个人。腾讯与这些公司之间签署独家协议，街景数据向搜搜独家供应。这些外部采集团队的规模约为“两三百人”，从上午 10 点到下午 4 点，一个采集车可以采集回来约 20GB 数据，这些车队一天总共可以采集回“几个 TB 数据”，一年加起来有一个 PB。

用户（个人或商户）还可以“邀拍”，搜搜街景地图团队依据用户的呼声来拍摄更精细的街景。另外，从实景采光效果来看，搜搜街景的 360 度照片确实非常透彻明亮，显然是经过刻意筛选的。SOSO 街景的高清景象可以帮助用户通过实景的方式更真实、快速地认识一个地点，其主要用途如表 14-5 所示。

表 14-5　SOSO 街景的主要用途

主 要 用 途	细 节 说 明
在线旅游	SOSO 街景可以提供 43 个城市和地区的街景，只要坐在电脑前就可以真实地看到街道上的高清景象，如图 14-6 所示
认清道路，快速到达目的地	去陌生的地方前，用户可以使用 SOSO 街景先提前看一看路况，使自己少走弯路
了解家人、朋友的生活环境	使用 SOSO 街景，可以让彼此看到居住的城市、街道，甚至可以看到你家的窗户。虽然相距千里，让彼此的心更亲近
买房租房，先用 SOSO 街景	买房租房的用户一定都吃过东奔西跑的苦头。利用 SOSO 街景，可以先看看你的目标小区长什么样子，周边环境如何……不但可以节省时间，结合 SOSO 地图丰富的查找功能，还可以坐在电脑前就轻松对比各个楼盘的周边环境

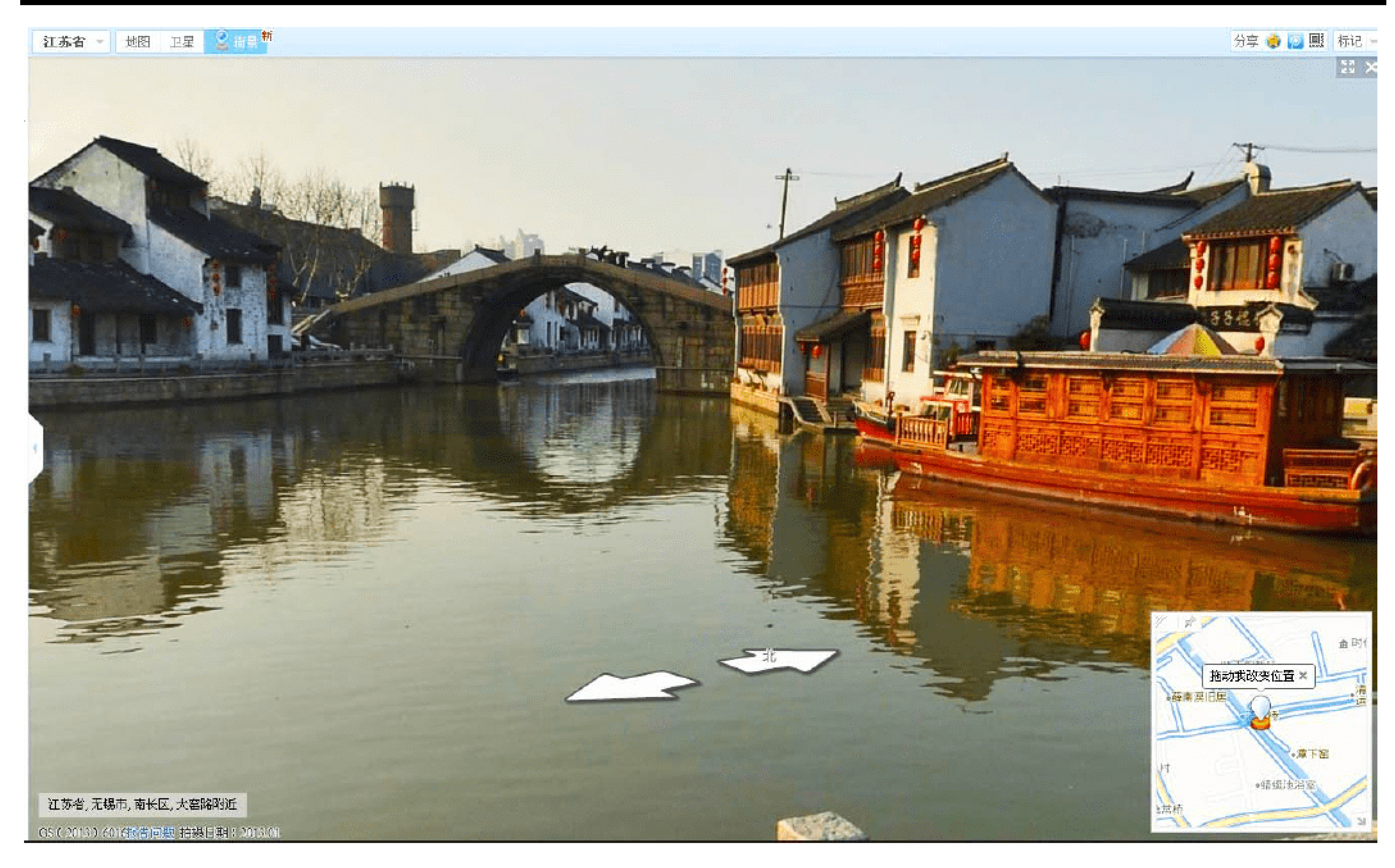


图 14-6　SOSO 街景地图

此外，SOSO 街景在用户体验上也进行了大量创新，如白天与夜景一键切换、图像清晰度提升、移动流畅度提升、画面惯性系统等。考虑到目前国内用户的网络带宽问题，SOSO 街景地图采用的图片经过压缩，但 SOSO 地图已经做好了提供更好视觉效果的准备，随着网络环境的改善，未来将推出更高画面质量的街景产品进行迭代。

当然，SOSO 街景地图也面临一定的挑战，笔者认为至少包括以下三个方面：

（1）街景地图管理政策。目前这个问题基本得到解决，腾讯凭借自己强大的关系网，成为街景地图监管政策的推动者之一，可谓因祸得福。

（2）天气和海量数据。为了获取更好的街景照片，数据收集人员只能“靠天吃饭”，这一点 Google 街景团队也概莫能外。

（3）数据量的存储。每年处理 1PB 数据对于腾讯来说是一个巨大的挑战，据悉腾讯公司已经为街景地图投入了“数亿元”人民币。

不久前，腾讯公司董事会主席兼首席执行官马化腾提出四大战略方向，包括 SoLoMo、照片、Voice 和手机安全，其中提到腾讯开放平台每天调用 LBS 数据的次数是 7 亿次，而且还在不断暴涨。街景地图的布局，如同腾讯 5 亿元影视投资基金一样，都属于长线投资。

【案例解析】：在本案例中，SOSO 街景地图的核心技术均采用自主技术，其中包括 3D 引擎、云平台存储计算及配套的图形图像技术。街景地图服务能够为 QQ 用户，尤其是年轻用户提供差异化的体验，增加 SOSO 地图的产品黏性，由此提升 SOSO 品牌影响力。另外，SOSO 街景地图的上线有利于推动国内在线地图产业的发展，触发行业的跟随效应，激发各家在线地图平台推出自己的新一代地图服务。

数据是街景地图的核心竞争力，既需要数据的数量，又要保证数据的质量。SOSO 地图首先保证了一条，即数据的独家性，这个优势的建立归功于起步早、行动快。地图产品本身意味着高成本的投入，加入了街景功能，意味着更大规模的支出。笔者认为，街景在未来还有大量的挖掘空间，包括更多的城市、更快的更新频次、形成历史变迁的时空记录、和 O2O 进行结合、用户个性化和社区化的相片分享等。

14.2.4 【案例】用大数据 APP 缓解交通压力

上海是一个人口和产业特别密集的特大城市，中心区 90 多平方公里之内平均每平方公里超过 4 万人，人均道路面积只有 2 平方米，只有国内同类城市的 1/2 到 1/3，国外同类城市的 1/5 到十几分之一。随着经济的发展，车辆增加很快，上海的道路交通负荷从总体上说已处于超饱和状态。这种交通的超饱和状态，不是采取一般的管理措施所能够解决的。而随着改革开放，城市的发展，这种矛盾还会迅速加剧。因此上海必须加快交通建设，锲而不舍地把解决城市交通问题作为城市建设的重点。

大数据在上海交通中已经有了广泛运用。上海从 2004 年开始，经过近十年的持续建设和应用，基本实现了对全市中心城区主要地面道路、城市快速路、高速公路信息采集和发布的覆盖。目前，对交通信息的采集主要是通过地磁线圈、出租车 GPS 信息、视频图像、信号控制系统等方式，采集车速、流量、交通事件等实时数据，经过网络传输汇聚到交通综合信息平台，实现跨部门交通数据的汇聚、共享与交换。

上海“智行者”APP 主要实现用户对上海市路况的整体了解，以简图的形式呈现给用户，方便用户及时掌握市内主要区域的道路状况，可以根据不同的路况优化行车路线，

节约旅行成本，如图 14-7 所示。当你驾车驶入指定区域时，会提前弹出该区域的交通路况简图，并对事件、施工、阻断等信息语音提示，使用户可以提前掌握该区域路况，随时变更行车路线。

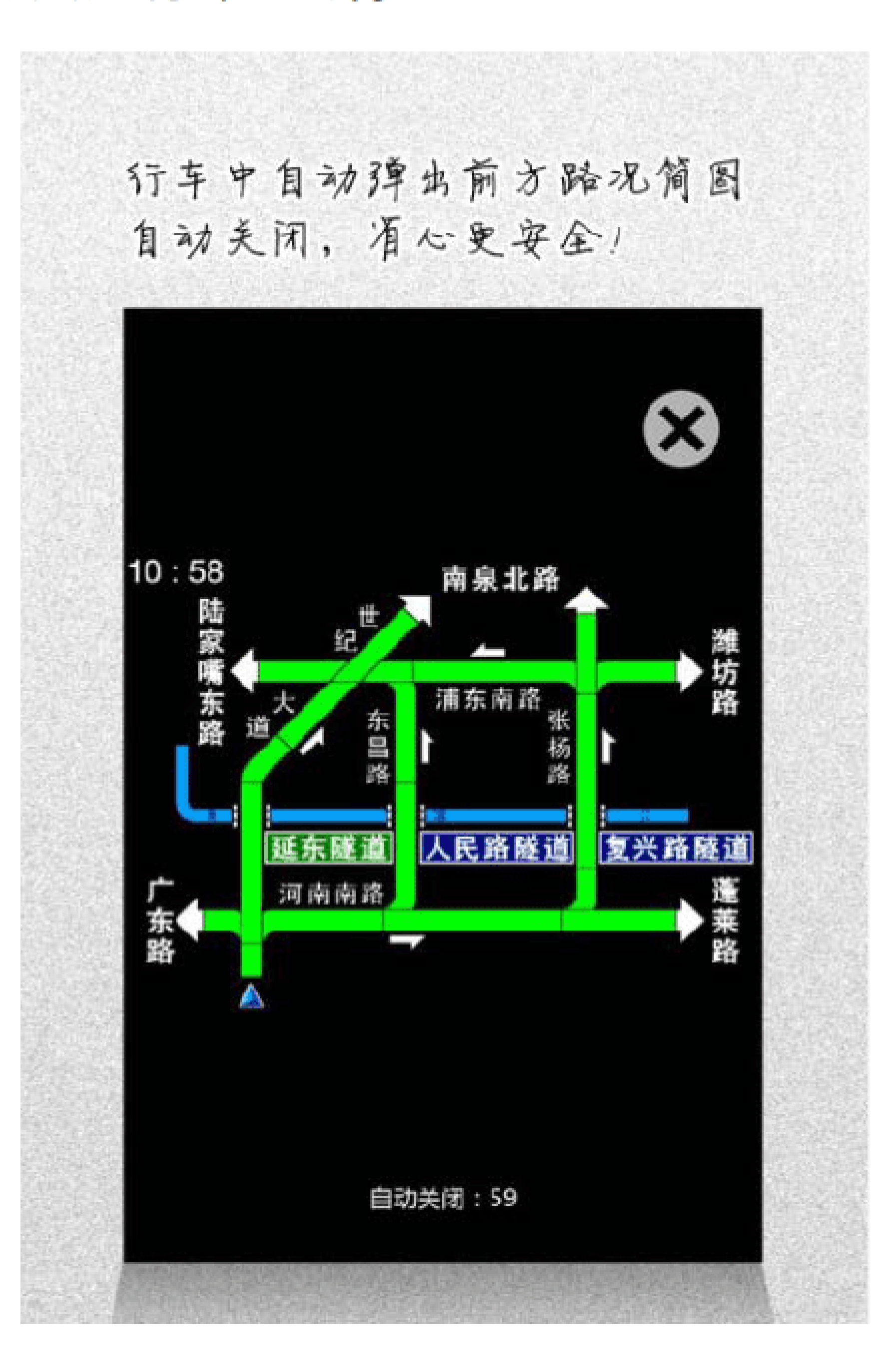




图 14-7　“智行者”APP 界面

针对不同路网的交通特征，通过获取包括数字、视频、图像等多种类型的交通数据，经数据的分析处理，获得道路交通通行指数和通行状态，通过车载终端、智能手机、网站和电台、电视等，多载体、多方式地实现交通状态信息的发布服务。

【案例解析】：在本案例中可以看出，大数据的分析和应用在上海道路交通中发挥了重要作用。笔者认为，智能交通技术可以有效地提高现有交通资源的使用效率，降低能耗，同时提高交通便捷水平和安全性，不同城市应据不同的规划和情况制定适合本地的智能出行方案。

因此，在驾驶者无法预知交通的拥堵可能性时，大数据亦可帮助用户预先了解。例如，在驾驶者出发前，大数据管理系统会依据前方路线中导致交通拥堵的各种因素，确定避开拥堵的备用路线，并通过智能手机告知驾驶者。

14.2.5　【案例】ETC 电子收费系统加大通行力

目前，全美公路总里程达到 630 多万公里，其中高速公路总里程已近 9 万公里。在高速公路的运营过程中，根据运营报表统计数据，人工半自动收费车道（Manual Ton

Collection System，MTC）的平均通行能力为 200 辆/小时，电子收费车道的平均通行能力为 1500 辆/小时，l 条 ETC（Electronic Toll Collection，即电子不停车收费系统，如图 14-8 所示）车道的通行能力是 MTC 车道通行能力的 7 倍。



图 14-8　ETC 收费通道

ETC 是目前世界上最先进的路桥收费系统，ETC 技术是以 IC 卡作为数据载体的，通过无线数据交换方式实现收费计算机与 IC 卡的远程数据存取功能。使用该系统，车主只要在车上安装 IC 卡并预存费用，通过收费站时便不用人工缴费，也无需停车，高速费将从卡中自动扣除，如图 14-9 所示。通过 ETC 系统，可以获取车主个人信息、卡内金额以及通行车速、时间、路径等。在数据获取方面，ETC 要远胜于摄像头监控、牌照识别、地感线圈等传统的车辆信息采集手段，采集到的信息也更加全面、准确。

美国最著名的联网运行电子不停车收费系统是 E-Zpass 系统，这种收费系统每车收费耗时不到两秒，其收费通道的通行能力是人工收费通道的 5～10 倍，在德国、日本、意大利都被广泛推广，其中意大利 30%的收费站安装使用了不停车收费设备，该收费方式每分钟平均可处理 30 辆车。

在美国，ETC 方式不但缓解了快速路、高速公路入口因人工缴费导致的拥堵情况，而且还成为美国回收公路投资和养护费用的高效率手段。另外，在海关和重要港口，使用 ETC 的车辆出了高速可以直接驶入码头，无需停车。ETC 在提高通行速度、减少拥堵、节能减排的同时，也为管理部门提供了出入车辆的基本数据。例如，用于对数据准确度和质量要求较高的监狱出入管理，通过分析每日车辆的进出记录，来核查是否存在非正常通行车辆。

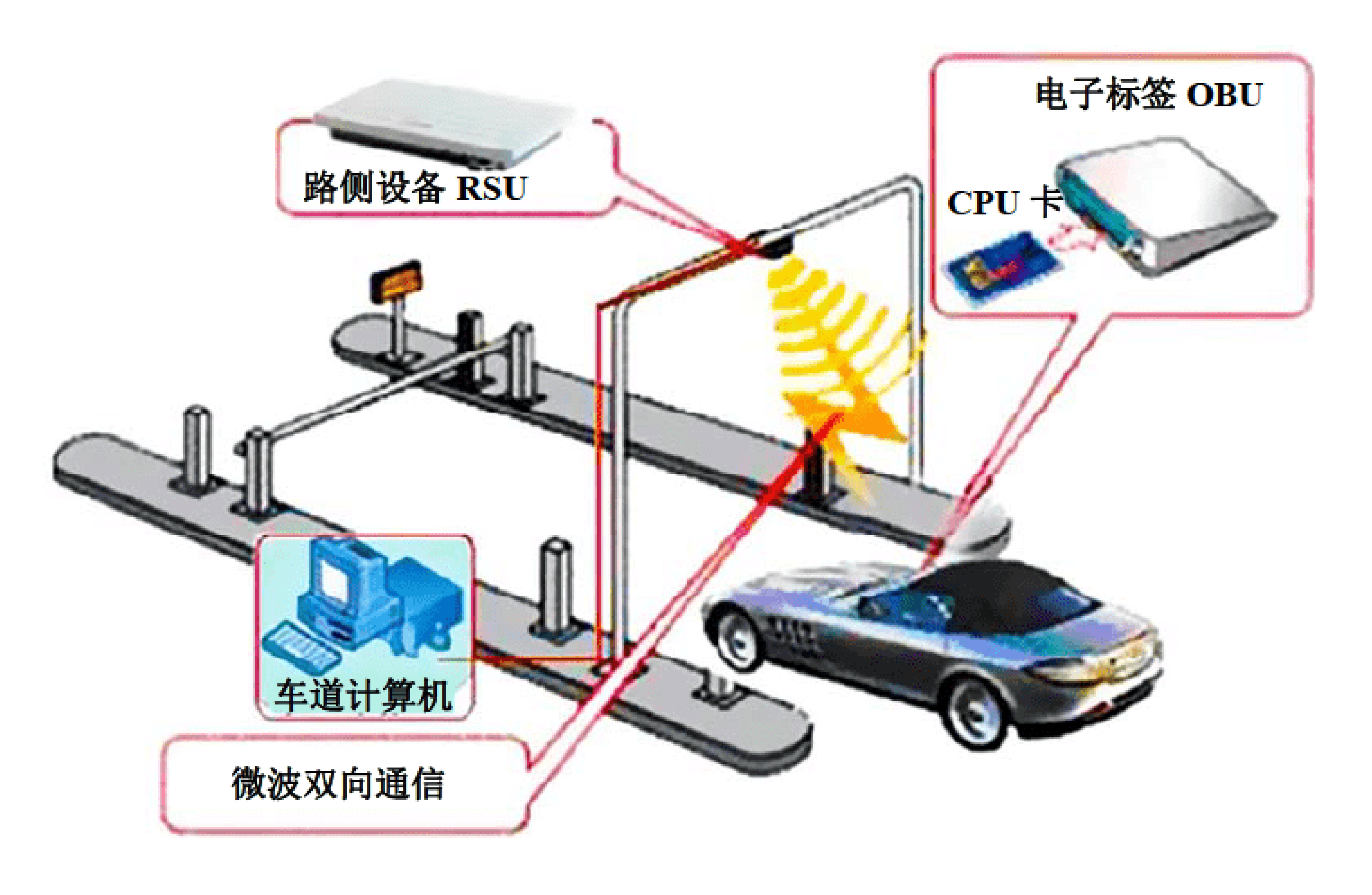


图 14-9　ETC 收费系统

【案例解析】：在本案例中，ETC 主要还是用于高速公路，其他扩展应用一方面是为了给用户带来更多便利，提供增值服务；另一方面，也便于政府加强监管，掌握更多管理数据。

基于 ETC 数据的收集原理，笔者认为，用户可以积极上报共享周边路况信息，为政府制定缓解城市交通拥堵决策提供依据；用户还可通过各种通信手段及时地将周边发生的交通状况和事件上报政府部门或相关企业，还可提出更为准确直接的交通缓解措施或方案。

总之，随着信息通信技术的发展，交通运输从数据贫乏的困境转向数据丰富的环境，而面对众多的交通数据，如何根据用户需求从中提取有效数据成为关键所在。

读书笔记



15

社会：用数据改变生活

学前提示

对于生活在社会中的普通人来说，大数据似乎离我们甚远，它看不见也摸不着，但又时时影响着人们的日常生活，那么人们在日常生活中有哪些事情涉及大数据呢？本章介绍大数据在教育、体育、影音媒体等生活中的应用案例，让你了解大数据到底改变了人们哪些生活方式。

要点展示

- 教育领域大数据应用案例
- 体育领域大数据应用案例
- 影音媒体大数据应用案例
- 生活中的大数据应用案例

15.1 教育领域大数据应用案例

大数据在社会诸多领域催生了很多变革，本节从教育领域探讨大数据的应用，并以此管窥大数据引发的重要变革。本节主要介绍大数据在教育领域的应用案例，希望对读者有一定的启发和学习价值。

15.1.1 【案例】大数据让在线教育变为现实

哈佛大学以及麻省理工学院在 2012 年联合发布了一款非营利性质的在线教育服务——edX。edX 平台在 2012 年还发布了课程编辑助手 Course Builder，其可以帮助教育机构编写自己的在线课程。

近日，谷歌也开始与 edX 合作，将强强联合推出 MOOC（Massive Open Online Course，巨多在线课堂，网址为 mooc.org）在线课堂。MOOC 将是一个面向于教育机构、政府、商业机构以及个人的在线教育平台，认证机构可以在 MOOC 上推出自己的课程，如图 15-1 所示。

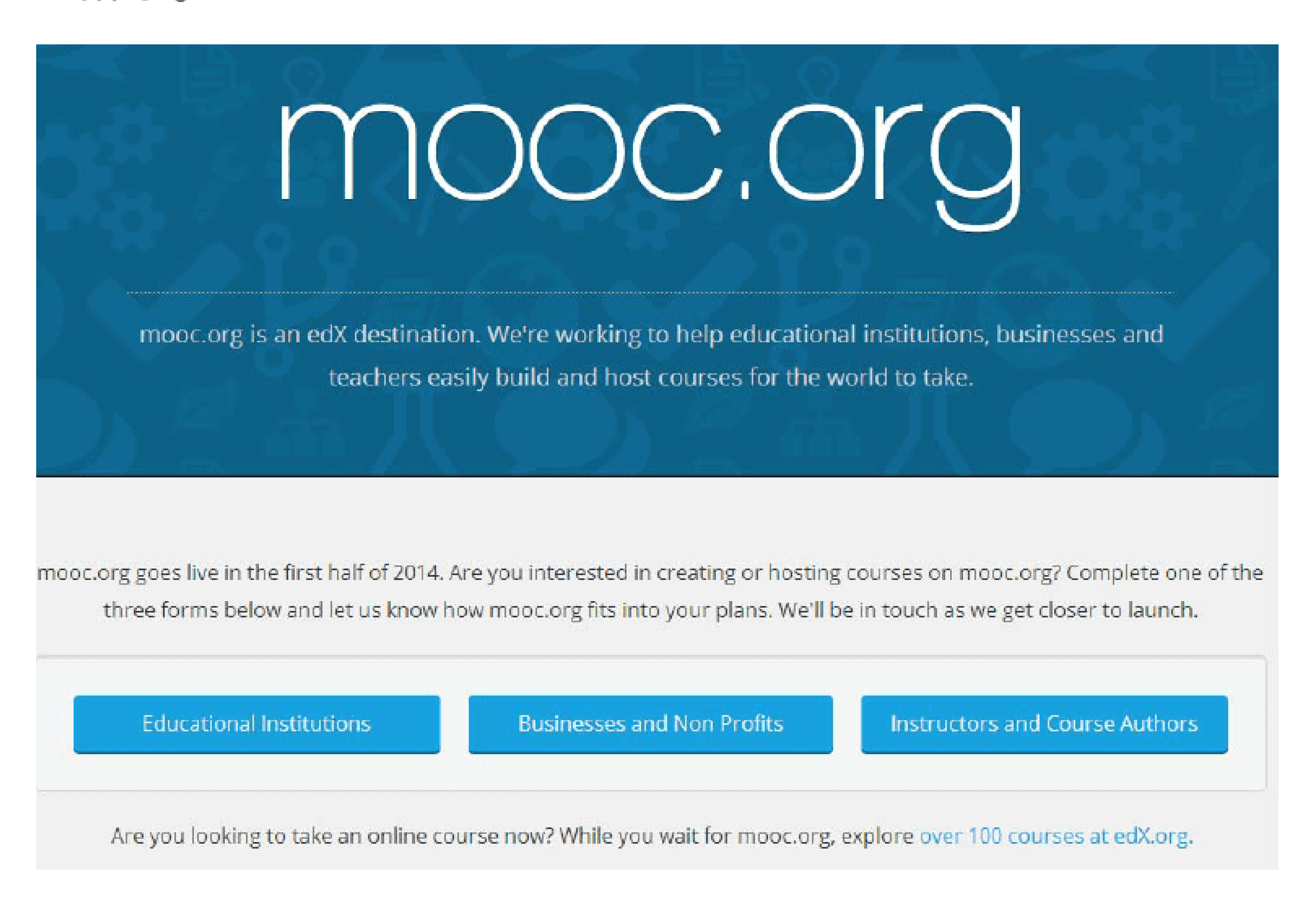


图 15-1　mooc.org 主页

到目前为止，edX.org 网站上的课程已经有 120 万名学生在使用。edX 提供的课程都是“受到管理的”，提供名牌大学的质量保证。与此相比，mooc.org 网站上的课程则将更具多样性，包括来自于公司和非营利机构的在线课程等。

【案例解析】：毋庸置疑，在国家大量需要科学、技术、工程和数学专业的毕业生之际，MOOC 是一项革命性的创新。假如你不是那种“文凭狂”，只想在比较好就业的

专业领域提升自我能力，MOOC 更可以说是一场教育革命。教育领域正在发生的这场革命，其深厚的技术背景就是由于信息技术的进步，人类收集、存储、分析、使用数据的能力实现了巨大跨越，这种现象也被称为“大数据”。

不难看出，未来的在线教育平台之所以强大，在于其能收集、分析、使用大量的数据。数据是对信息的记录，数据的激增意味着人类的记录范围、测量范围和分析范围在不断扩大，也意味着知识的边界在不断延伸。大数据将对人类社会发生的影响难以限量，以行为评价和学习诱导为特点的在线教育平台只是这个大潮在教育领域掀起的一朵浪花。

15.1.2 【案例】无孔不入的数字化学习平台

日本网络大学(Cyber University)是一所位于日本福冈县的公司式经营的私立大学，是日本唯一的在互联网提供全部课程的大学。网络大学原来面向网络用户提供课程，这些课程的内容包含图片、视频以及声音，而手机版的课程为 PowerPoint 图片流媒体视频。

例如，网络大学在手机上提供一节“金字塔的秘密”的课程，金字塔的图像出现在手机屏幕上，然后，从手机的扬声器中播放出教授的声音，而且图片也会根据语音内容不断地变换。

据悉，网络大学预期将在手机上提供大约 100 种课程，其中包括中国文化、在线新闻和英国文学。与其他课程不同的是，用手机向公众讲课是免费的，但观众需要支付手机费用。

在网络大学的规定中，学生们要通过宽带互联网上课，并且向教授上交自己的作业论文。在完成所有课程和论文之后，学生可以得到正式的本科学历。

专 家 提 醒

实际上相似性质的网络大学也曾经在其他国家出现，例如，美国 Phoenix 大学，建立于 20 世纪 70 年代，目前已经在北美地区招募到超过两万名学生，它的绝大部分课程都通过网络形式教授。

【案例解析】：在本案例中，网络大学为那些无法上实体大学的人提供了受教育机会，尤其是上班族、残疾人和病人。

其实，笔者认为网络大学还可以结合流行的大数据技术，利用流媒体视频和数据分析帮助教师跟踪学生的学习情况，根据他们的能力水平定制教学内容，以及预测学生的执行情况。

15.1.3 【案例】用云平台全面推进素质教育

亚洲教育网自主研发的“三网智慧泛教育云平台”，为国内教育部门和学校构建了支持“三网融合、泛在学习”的公共智慧云，形成“学校—家庭—社会”三位一体的绿

色网络平台。

据悉，该平台全面支持素质教育和绿色评价体系，以开放共享的“公共云”消除地区和学校的信息孤岛，以电脑、手机、电视、平板等多终端实现了教师、学生、家长的轻松访问，让先进的教育理念和优质的教育资源可以覆盖到农村和偏远地区，从而有力地促进了教育公平和教育均衡发展。

“三网智慧泛教育云”利用云计算、物联网和虚拟化等新技术来升级校园网、城域网，其创建的“教育云＋互动电视＋电子书包”新模式开启了教育信息化新纪元。

“三网智慧泛教育云”包含互动社区云、教育管理云及教学资源云三大子云。

- 互动社区云。为学校、家庭和社会之间的多向互动交流提供开放共享的信息交互平台，用户只需一个账号就能实现多个平台间的访问和多重交流。
- 教育管理云。帮助学校整体规划教育信息化应用，聚合学校各管理事务所需的子系统，支持学校按需拓展及升级应用系统，促进学校低成本实现校园数字化管理。
- 教学资源云。使分散、异构的学习资源能够进行有机整合，从而促进教学资源的优化配置。

“三网智慧泛教育云”全面推动教育信息化，目前亚洲教育网正逐步地将教育云平台与物联网进行高度融合，以方便用户灵活接入各种软硬件系统，力求最终全面实现“学习交流人人通、资源共享班班通、优质教育校校通”的教育信息化整体解决方案，全面推进素质教育。

【案例解析】：云教育是指基于云计算商业模式应用的教育平台服务。在云平台上，所有的教育机构、培训机构、招生服务机构、宣传机构、行业协会、管理机构、行业媒体、法律机构等都集中云整合成资源池，各个资源相互展示和互动，按需交流，达成意向，从而降低教育成本，提高效率。

在本案例中，可以看到云计算技术在高校的发展，已经从原来的理论步入实际应用。基于大数据的云可以用来共享教育资源、分享教育成果，使教育者和受教育者实现互动，如图 15-2 所示。

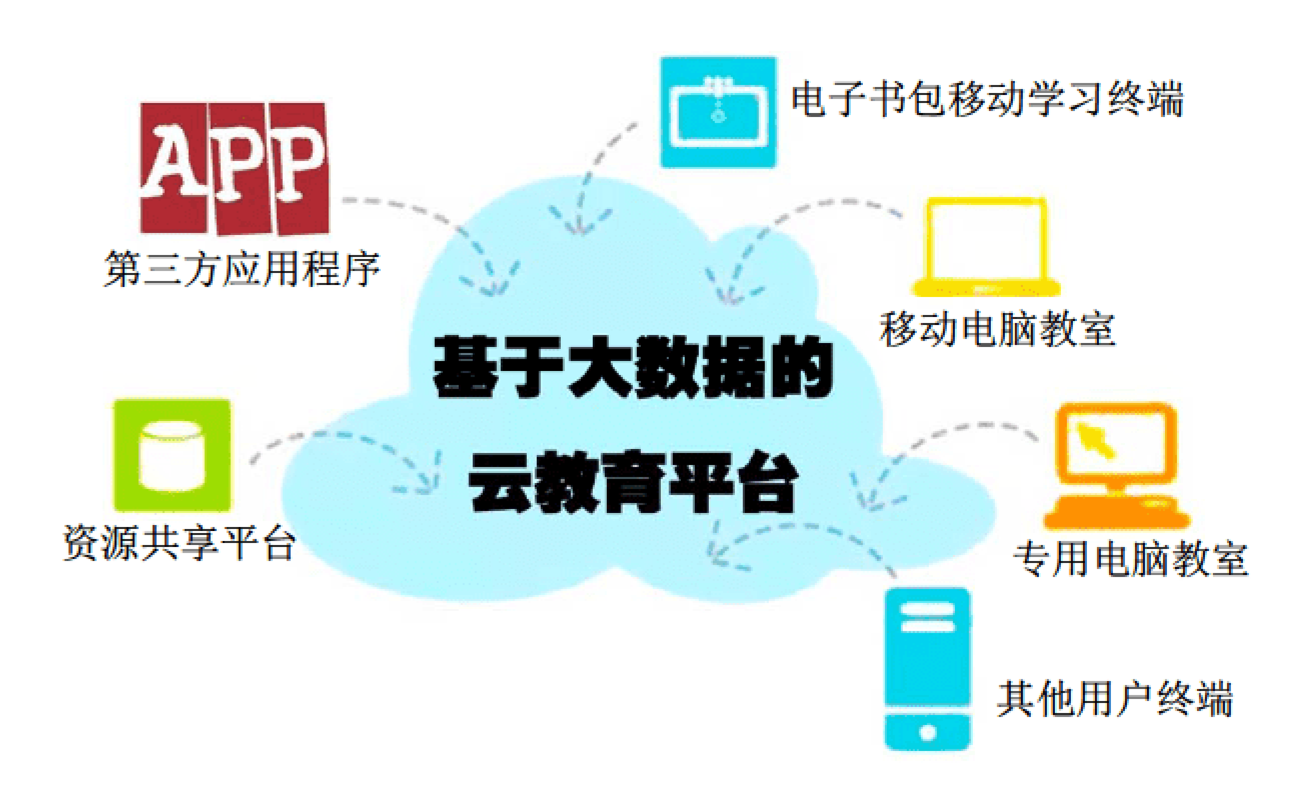


图 15-2　大数据云教育平台的功能

专家提醒

如果说大数据本身是一个问题集，需要一个“管家”来处理。那么，“云技术”就是解决大数据问题集最重要、最有效的手段。

15.1.4 【案例】美国政府用大数据改善教育

近年来，美国高中生和大学生的教育情况不容乐观：高中生退学率高达 30%（平均每 26 秒就有一个高中生退学），33%的大学生需要重修，46%的大学生无法正常毕业。对此，美国联邦政府教育部 2012 年参与了一项耗资两亿美元的有关公共教育的大数据计划，该计划的目的是通过运用大数据分析来改善教育。联邦教育部从财政预算中支出 2500 万美元，用于理解学生在个性化层面是怎样学习的。

美国教育部门运用大数据创造了“学习分析系统”，它是一个数据挖掘、模块化和案例运用的联合框架，可以向教育工作者提供了解学生到底是在“怎样”学习的更多、更好、更精确的信息。

例如，一个学生成绩不好是由于他因为周围环境而分心了吗？期末考试不及格是否意味着该学生并没有完全掌握这一学期的学习内容，还是因为他请了很多病假的缘故？利用大数据的学习分析能够向教育工作者提供有用的信息，从而帮助其回答这些不太好回答的现实问题。

【案例解析】：在本案例中，“学习分析系统”可以通过大数据技术，允许中小学和大学分析从学生的学习行为、考试分数到职业规划等所有重要的信息。许多这样的数据已经被诸如美国国家教育统计中心之类的政府机构储存起来用于统计和分析。

如今，互动性学习的新方法已经通过智力辅导系统、刺激与激励机制、教育性的游戏产生了越来越多的尚未结构化的数据。因此，笔者认为，教育中的非结构化数据（Unstructured Data）挖掘是迈向大数据分析的一项主要工作，更丰富的数据能给研究者提供比过去更多的探究学生学习环境的新机会。

15.1.5 【案例】大数据有效地指导学生学习

“渴望学习”（Desire 2 Learn）是一家总部位于加拿大安大略省沃特卢的教育科技公司，其推出了基于他们自己过去的学习成绩数据预测并改善其未来学习成绩的大数据服务项目。

Desire 2 Learn 公司的新产品名为“学生成功系统”（Student Success System），该产品通过监控学生阅读电子化的课程材料、提交电子版的作业、在线与同学交流、完成考试与测验，就能让其计算程序持续、系统地分析每个学生的教育数据。

利用“学生成功系统”，老师得到的不再是过去那种只展示学生分数与作业的结果，

而是像阅读材料的时间长短等这样更为详细的重要信息。因此，老师可以及时诊断问题的所在，提出改进的建议，并预测学生的期末考试成绩。

据悉，加拿大和美国的 1000 多万名高校学生正在使用“学生成功系统”来改善学习成绩。

【案例解析】：在本案例中，Desire 2 Learn 公司通过大数据创建的学习分析系统，可以有效地指导学生朝着更加个性化的学习进程迈进。

在大数据时代，通过大数据进行学习分析能够为每一位学生都创设一个量身定做的学习环境和个性化的课程，还能创建一个早期预警系统以便发现开除和辍学等潜在的风险，为学生的多年学习提供一个富有挑战性而非逐渐厌倦的学习计划。

专家提醒

大数据与传统数据的区别在于人们对于“数据”的理解更为深入了，许多我们曾经并没有重视的，或者缺乏技术与方法去收集的信息，现在都可以作为“数据”进行记录与分析了。

15.1.6 【案例】用大数据管理上海大学招生

在“大数据”概念未出现时，上大已经开启了数据信息库的“基础设施建设”。从 1998 年开始，当时作为学校“招生官”的叶志明就要求行政部门工作人员勤录数据、筹建信息库。所有的数据都要按照规定的格式录入，并同时设定不同数据的属性，在当时，这被认为是“繁琐得要命”的事情，并不讨好。但时至今日，海量数据已对上大的教育教学管理和改革发挥了非常积极的作用。

2012 年，上海大学宣布退出春季高考（以下简称春考）。业内有分析说，除了报考人数下降外，春考给学校日常教学管理带来难题，甚至考务成本高昂等，是大学对春考“不感冒”的原因。

此时，已经是上海大学副校长的叶志明表示：“同样是探索打破传统高考制度的新举措，上大决定退出春考，但今后会更加支持插班生考试。这些决策的依据，正是一揽子和这两项招生考试相关的数据。”

上海从 2000 年率先推出春季高考。同年，上大招收插班生的人数为 55 人，到 2011 年时，插班生招生数达 152 人。和秋季高考进校的学生作比较，统计数据表明，插班生的学习情况，历年来都优于秋考生。但同期通过春考招收的学生，除了 2001—2004 年的平均成绩超过秋考生外，往后的年份里，春考生表现一路走低，2009 年时，春考生的平均成绩更是被秋考生甩开了一大截。

学校通过分析近 10 年的招生数据，很快找到了其中的原因。2008 年以后上海高考招生实行平行志愿，考生由于填志愿等原因落档继而选择复读的人数锐减。眼见春考生源一年不如一年，上大决定退出，把招生名额用于生源更佳的插班生考试。

上大还用数据来处理延期毕业的学生。以上大 2008 级学生为例，申请延期的有 580

人。统计表明，其中超过七成是因为大一、大二时的公共基础课和专业基础课“挂科”。另外，学生最容易不及格的课程依次是高等数学、大学物理、概率论与数理统计、大学英语以及计算机等。通过在学生的数据库里搜索和分析相关数据，就可以轻而易举地找到学校里挂科率较高的学院。然后，通过这个数据库，将延期问题进行聚焦，发现很多学生无法如期毕业，隐患在大一、大二时就已经埋下了。因此，辅导员和院系分管教务工作的老师可以多关照大一、大二学生的基础课，将学习盯得紧一点，即可解决相关的延期 问题。

【案例解析】：在本案例中，上海大学是一所面向全国二十多个省市招生的高校，需要的生源在哪里，应该向哪些省份多投放招生名额，这些具体决策需要数据支撑。因此，笔者觉得上大已经在大数据战略上迈出了重要的一步，今后还可以从其他方面继续努力，让大数据管理支持更多的学习决策。

随着数据越积越多，高校人士也开始意识到，这些数据会“说话”，能在办学中派更大的用场。笔者认为，“沉睡多年”的教育数据已经苏醒，大数据参与学校的教育教学管理尤其是改革方向的决策，上海大学只是这其中的一个样本，但足以让人看到一个事实：大数据时代，高校教育也正由此可以发生变革。

15.2 体育领域大数据应用案例

尽管科学家预言大数据将改变未来人类生活的方方面面，但它确实首先在体育赛事中展现了自己的价值，并彻底颠覆了传统的体育理念。本节主要介绍大数据在体育领域的应用案例，希望对读者有一定的启发和学习价值。

15.2.1 【案例】Nike 记录运动中的数据价值

Nike 作为全球最大的运动品牌公司之一，曾在官网上公布了这样两则信息：“在冬天，美国人比欧洲和非洲人都更喜欢跑步这项运动，但美国人平均每次跑步的长度和时间都比欧洲人短”，所以 Nike 计划在不同的市场区域做好不同的产品划分，运动鞋的设计也根据区域的不同做了独立调整。

耐克公司与苹果电脑公司这两家全球首屈一指的大型公司终于推出了合作后的第一款产品 Nike Plus，它可以让耐克公司的运动鞋和苹果电脑公司的 iPod Nano 便携式媒体播放器进行通信。Nike + iPod 运动联合系统包含一个放置在耐克运动鞋衬垫下的小巧的椭圆形晶片（有点类似 SIM 卡）和一个装备在 iPod Nano 便携式媒体播放器上的小型传感器，如图 15-3 所示。



图 15-3　Nike+iPod 运动联合系统

Nike Plus 相关的软件除了可以捕获像时间和距离这样的一般数据外，其还包含有一个语音系统可以交流更多的信息，这有点类似汽车上的导航系统。另外，Nike Plus 还可以给运动者提供运动的激情，耐克公司搜罗了兰斯・阿姆斯特朗和保拉・拉德克利夫在运动时的一些心得体会，后者是马拉松纪录的保持者。这样我们在运动时就可以分享这些运动大师最喜爱的音乐和运动激情所在了。

苹果电脑公司的 iTunes 音乐在线零售商店也增设了一个耐克运动音乐区域为喜爱运动的消费者提供体能测验以及运动激情等。

消费者在进行运动测验时，iPod Nano 便携式媒体播放器的屏幕上可以显示相关的测验数据以及测验总结等。iPod Nano 便携式媒体播放器上显示的自己体能测验数据可以上传到 nikeplus.com 网站上，如图 15-4 所示。

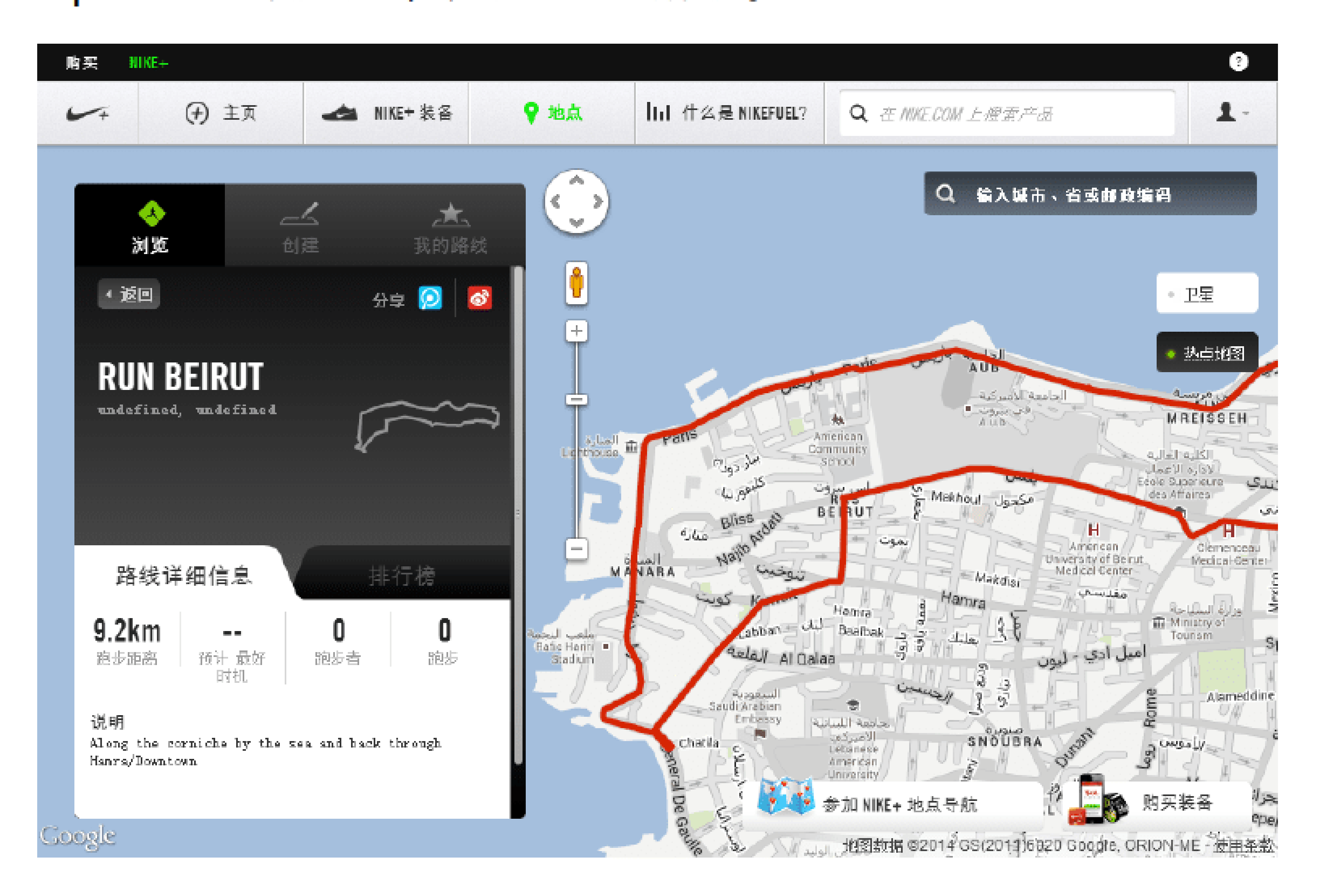


图 15-4　nikeplus.com 网站

nikeplus.com 上有实时数据更新，因此使用者对自己跑步的公里数、消耗的卡路里

以及路径都能了如指掌，还可以分享并关注朋友们取得的进步，这个创新不仅仅使 Nike Plus 变成了体育运动爱好者的 Facebook，Nike 也成功建立了全球最大的运动相关的网上社区（有超过 5 百万的活跃注册用户，上传超过几十亿公里数和几百亿卡路里数）。

【案例解析】：在本案例中，Nike 的成功和市场上的特立独行正是来源于对自身产品和消费者的数据挖掘。

试想一下，如果一双专业跑步鞋除了给人们提供足够的运动性能以外，同时又要适合各种运动员的穿着与跑步，那么没有一个跑步数据测试工具，怎么能够测试出运动员要怎么跑才能减少失误与提高效率呢？因此，如果在一双耐克跑步鞋上装上 Nike Plus 跑步数据工具，就能更快、更准确地测出运动员跑步的效率，以及了解自己要怎么跑才能够提高效率。

15.2.2 【案例】大数据助力 NBA 赛事全过程

NBA（National Basketball Association，即美国篮球职业联赛）早从 1980 年就开始使用数据管理技术，统计所有球员得分、篮板、助攻、盖帽、抢断、失误、犯规等一系列场上数据，如图 15-5 所示。NBA 通过详实而细致的数据统计，不仅可以提供单个球员的查询服务，还可以对比两名球员，包括两人对位攻防时的表现，并进行数据化分析。例如，詹姆斯场均能得 28 分，科比得 27 分，但当两人相遇时，科比场均能得到 30 分，詹姆斯只有 24 分。

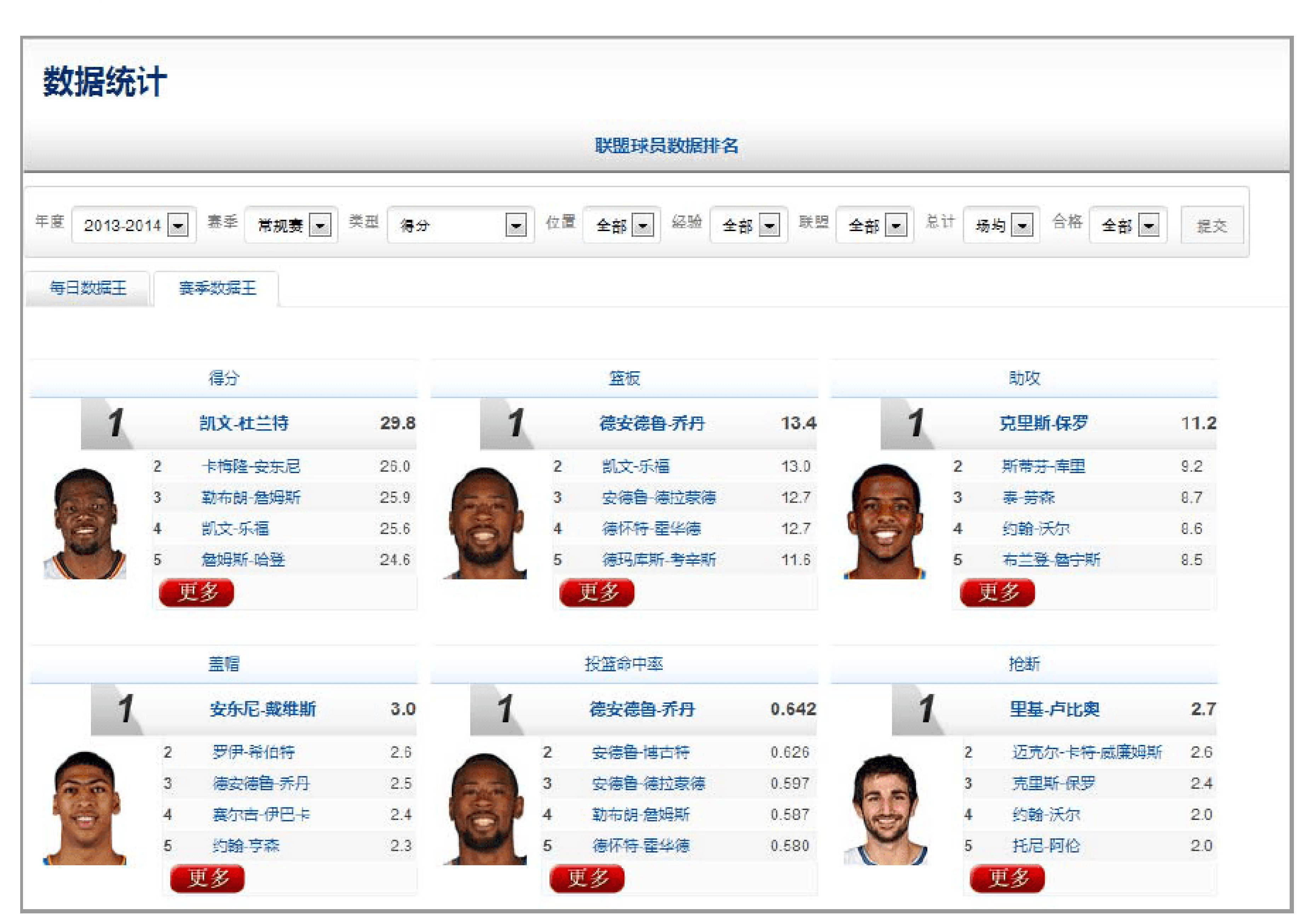


图 15-5　NBA 官网的球员数据统计

如今，NBA 的数据统计和管理更为成熟丰富，还能提供包括场上效率、得分区域等分析。例如，2012 年席卷 NBA 的华裔运动员林书豪，在爆发期间一直被专家诟病的一点就是失误太多。这正是来自强大的数据统计，他的助攻失误比仅为 2.0，也就是说每送出两个助攻就要伴随一次失误，而顶级后卫保罗的助攻失误比为 4.6，超出林书豪一倍，显然更为出色。

在 NBA 的中文官方网站上，有专门的统计页面，上面把 NBA 历史上收集的几乎所有球员、球队信息以非常易用的方式提供出来，后台使用了 SAP HANA 这样的内存分析数据库，以应对网站数以万计的访问者的访问，提高随机、灵活查询的速度，它提供了一种前所未有的用户体验，以及对上百个指标的不同过滤、统计和排序等，用户可以定制分析报表，而不需要大量固化报表格式和场景，如图 15-6 所示。



图 15-6　NBA 热区数据分析

【案例解析】：在本案例中，NBA 非常聪明，把这些数据开放出来，让大家都对它们感兴趣，让每个球迷都有可能“如数家珍”，增加球迷们对球星们的迷恋程度，也从而增加对 NBA 比赛的热爱程度。

一个看似并不“高科技”的体育项目，都可以如此利用“大数据”的手段，以提供非常优秀的用户体验，从数据收集到数据统计和挖掘，到优秀的数据展现，非常值得其他企业学习。有了这样严格、精细的量化，就有了科学的态度，也就有了科学的指导思想和手段。

15.2.3 【案例】大数据颠覆网球的游戏规则

到目前（2013 年）为止，IBM 与法国网球协会合作有 28 年了，为法国网球公开赛（以下简称“法网”）提供支持。IBM 为法网带来一系列解决方案，全部都以实时及历史大满贯赛事数据为中心。IBM 负责获取、分析、保护、存储和分发法网的全部数据，实际上，大数据是 IBM 与法国网球协会合作的核心。

IBM 以多种方式使用大数据改善网球比赛，将法网的行动带给世界各地的球迷、教练、球员和媒体。例如，使用 SlamTracker 分析工具改变了许多球迷观赏网球比赛的方式。

SlamTracker 分析 8 年的法网网球比赛数据（每场比赛 4100 万个数据点），为每个球员确定将影响一场特定比赛的三项关键策略，并将其称之为“比赛的关键点”（keys to the match）。在比赛前，球迷可登录网站查看每个球员在一场比赛中的关键点，在比赛期间根据这些关键点，逐项实时观看球员的进步，如图 15-7 所示。



图 15-7　查看每个球员在比赛中的关键点

【案例解析】：在本案例中，通过 IBM 的 SlamTracker 数据分析工具，系统可以从过去的激烈比赛中过滤并且排列每位选手在比赛中的三个最重要的得分。例如，一个选手第二次发球可能需要达到一定比例才能获胜，或者长球得分是否有利于某位对手。在比赛之前了解关键进球，然后在比赛进行过程中关注选手的表现，用户可以实时看到关键进球是成功的良好预测指标。

其实，这项技术不仅限于在体育比赛中应用，同样的分析软件也在医院用于监控产前病房的婴儿、在警察局用于预防犯罪，并且在金融服务公司用于改善客户服务并降低

成本。

15.2.4 【案例】从大数据中获得宝贵洞察力

IBM 作为温布尔顿网球锦标赛的赞助商，不久前向中心球场推出了一项名为 IBM SecondSight 的新技术。

IBM SecondSight 的想法来自两年前锦标赛的一个重大事件，当时，美国的 John Isner 和法国的 Nicolas Mahut 进行了一场专业网球比赛，这是历史上最长的一次比赛。183 局的比赛长达 11 个小时零五分钟，历时三天。期间，平局比赛的分数不断升高，计分系统的设计人员没有预测到需要记录并显示如此高的分数，面临着数字用完的风险。最后，Isner 以一记“超身球”结束了比赛，在平局比赛中获胜。

IBM 英国公司客户与计划业务主管 Alan Flack 从这次比赛中得到启发：“我们为何不追踪球员的运动？毕竟，我们记录了比赛的其他所有内容。”于是，Alan Flack 决定与一家主营业务是追踪导弹的技术合作伙伴共同开发这样的系统。

IBM SecondSight 借助多个按战略角度分布的摄像头采集数据，可以实时追踪球员的运动，并以数字化屏幕显示方式展现给球迷，并且带有表示球员的头像。球迷可以点击图标查看最新的比赛分析。谁的动作更快？谁跑得更远？是否有人累了？

【案例解析】：在本案例中，IBM SecondSight 展示了从比赛纯物理角度来讲最深层的视图，丰富了球迷（以及教练和官员）的网球知识。虽然处于初级阶段，但笔者能够想象到运动追踪技术在网球和其他体育比赛之外的领域中的强大用途。例如，这项技术可用于监控和分析商场、工厂、机场的人员移动，或者高速公路的车流，我们能够从这类信息中获得宝贵的洞察力。

15.2.5 【案例】用预测分析软件来防止受伤

在超级联赛十五人橄榄球赛中，莱斯特老虎队已经开始利用 IBM 的预测分析软件，来评估球员受伤的可能性，为处于险境的球员设计个性化的训练计划。

几个赛季以来，莱斯特老虎队一直在收集数据，以期获得竞争优势。莱斯特老虎队的数据收集几乎是不间断的，队员配备 GPS 监视器和加速器，这些设备测评他们的碰撞强度，同时收集数据来监控球员的疲劳程度，这是一项关键的伤害预测变量。常规的调查问卷也收集主观性的生活方式信息。

莱斯特老虎队的运动科学主管 Andrew Shelton 表示：“任何人都可以收集数据，但重要的是，如何利用这些重要的数据。我们希望能够更好地利用我们的数据，尽可能好地为每个球员提供最佳的表现机会。如果你在球场上有最优秀的球员，那么失利的可能性就小。这不是多么高深复杂的事情，我们想要向下挖掘数据，确定如何能防止球员受

伤。”

通过利用 IBM 的大数据预测分析软件，Shelton 的队伍可以看到一个球员的一项或多项疲劳参数是否发生了重大变化，因此如果球员要参加一个高强度训练项目，分析软件可预测重大伤害风险，球队可通过这样的洞察力相应改变个人的训练计划。

【案例解析】：使用数据使体育俱乐部能够更加科学地评价球员，这是另一个新兴领域。在本案例中可以看出，从挑选最高效的球员，到最大限度地减少受伤概率，以及改善球迷体验等，数据分析在体育世界的应用越来越广泛。

专家提醒

美国奥克兰市运动家棒球队，曾因采用数学模型来预测球员成绩、遴选球员而大幅改变了球队成绩，创造了美国棒球联赛史上最长的连续获胜纪录。此后，越来越多的球队开始运用预测模型评估球员的潜力和市场价值，而那些先行一步的球队几乎都赢得了显著的竞争优势，明显胜过比他们更保守的同行。

15.2.6 【案例】普通球迷也能成为分析专家

2012 年，腾讯网正式推出国内首创的 NBA 数据库大师，结合视频和专家分析，给球迷带来了全新的视频体验以及更真实、更全面的篮球享受，如图 15-8 所示。

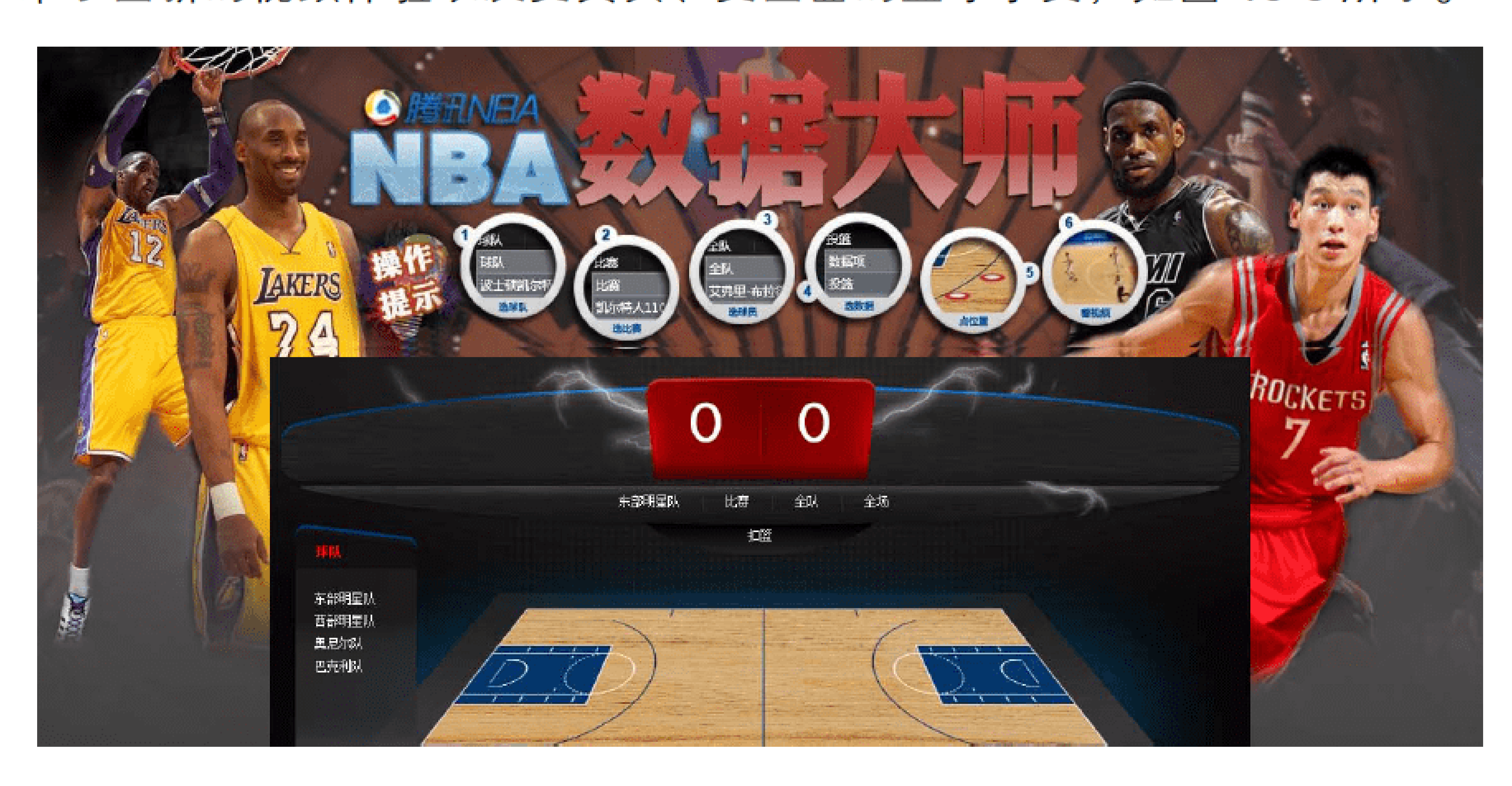


图 15-8 腾讯 NBA 数据库大师主页

NBA 数据库大师全面记录了 NBA 球星投篮、失误、助攻、抢断和犯规等 10 大数据，同时结合了比赛视频和专家解说分析的多功能数据系统。也就是说，NBA 数据库大师将会记录下每个球星在球场上的表现，包括投篮、三分、失误等。另外，球星的每个动作，都会用圆圈或者是箭头来表示，并且每个图标都含有视频的链接。用户随便点击一个图标，就会出现相对应的视频。

更重要的是，在该平台上，国内顶尖 NBA 专家还通过腾讯微博，随时随地与网友

分享自己的 NBA 见解。普通球迷在欣赏 NBA 专家点评的同时，可以与 NBA 大师进行讨论、互动，提高自己的水平，最终成为 NBA 分析大师。

【案例解析】：在本案例中，作为当前国内流量最高、影响力最大、产品线最全的门户网站腾讯，在 NBA 数据库大师平台上提供了海量的视频数据、独家的 NBA 资讯、国内顶尖 NBA 专家的专业解读，不仅给球迷带来了全新的视频体验以及更真实、更全面的篮球享受，并且帮助球迷成长为 NBAMaster（NBA 大师）。

笔者认为这是腾讯“点石成金”的关键一招，“大数据”加上“分析”，才有可能有价值，才有意义。“分析”才是关键能力，没有“分析”的“大数据”，就是一场淹没一切的数据海啸，是灾难。

我们可以用“分析”从大量的数据中寻找相关性模式，发现以前不为人知的、超越于平凡知识之上的、至关重要的新知识。这样的新知识，是隐藏在表象之下的获胜关键，是决定竞争结局的密码，是价值和财富。很多商业界的有识之士正是发现了这一点，才会狂热地追捧大数据。我们也可以想象一下，如果把这样的能力放在商业里，放在公共服务里，放在日常的工作和生活里，能给我们带来什么?

15.3 影音媒体大数据应用案例

经过两年的积淀与发展，新媒体影视业在 2013 年呈现爆发性增长。凭借对用户的精准定位，以及对市场的迅速反应，新媒体影视正在对传统影视形成极大冲击。笔者认为，精准的数据分析，将成为新媒体影视能否获得成功的关键。本节主要介绍大数据在媒体影视业的应用案例，希望对读者有一定的启发和学习价值。

15.3.1 【案例】《爸爸去哪儿》成口碑之王

眼下最炙手可热的真人秀栏目《爸爸去哪儿》，是中国湖南卫视从韩国 MBC 电视台引进的亲子户外真人秀节目，概念参考自韩国 MBC 电视台节目《爸爸！我们去哪儿？》；这是继湖南卫视《变形计》之后又一档真人秀亲子交互节目。

《爸爸去哪儿》讲述了 5 位明星爸爸跟子女 72 小时的乡村体验，爸爸单独肩负起照顾孩子饮食起居的责任，节目组设置一系列由父子（女）共同完成的任务，父子（女）俩在不熟悉的环境下状况百出。毫无疑问，亲子类的节目概念在中国电视圈内颇具创新意义。面对“父爱”普遍缺失的现状，湖南卫视的这档节目可以说是十分及时，不仅让爱回归，同样也能让初为父母的普通年轻人对育儿有一个全新的认识。

新华社新媒体中心联合数托邦工作室抓取了新浪微博上提及《爸爸去哪儿》的 45.5 万条原创微博，并对 36.7 万独立原发作者用户（去除疑似水军账户）、1300 余万条用户

微博及近 1 亿的关系进行数据分析，如图 15-9 所示。《爸爸去哪儿》不仅成为名副其实的“口碑王”，还使娱乐节目发生了很多微妙的变化。

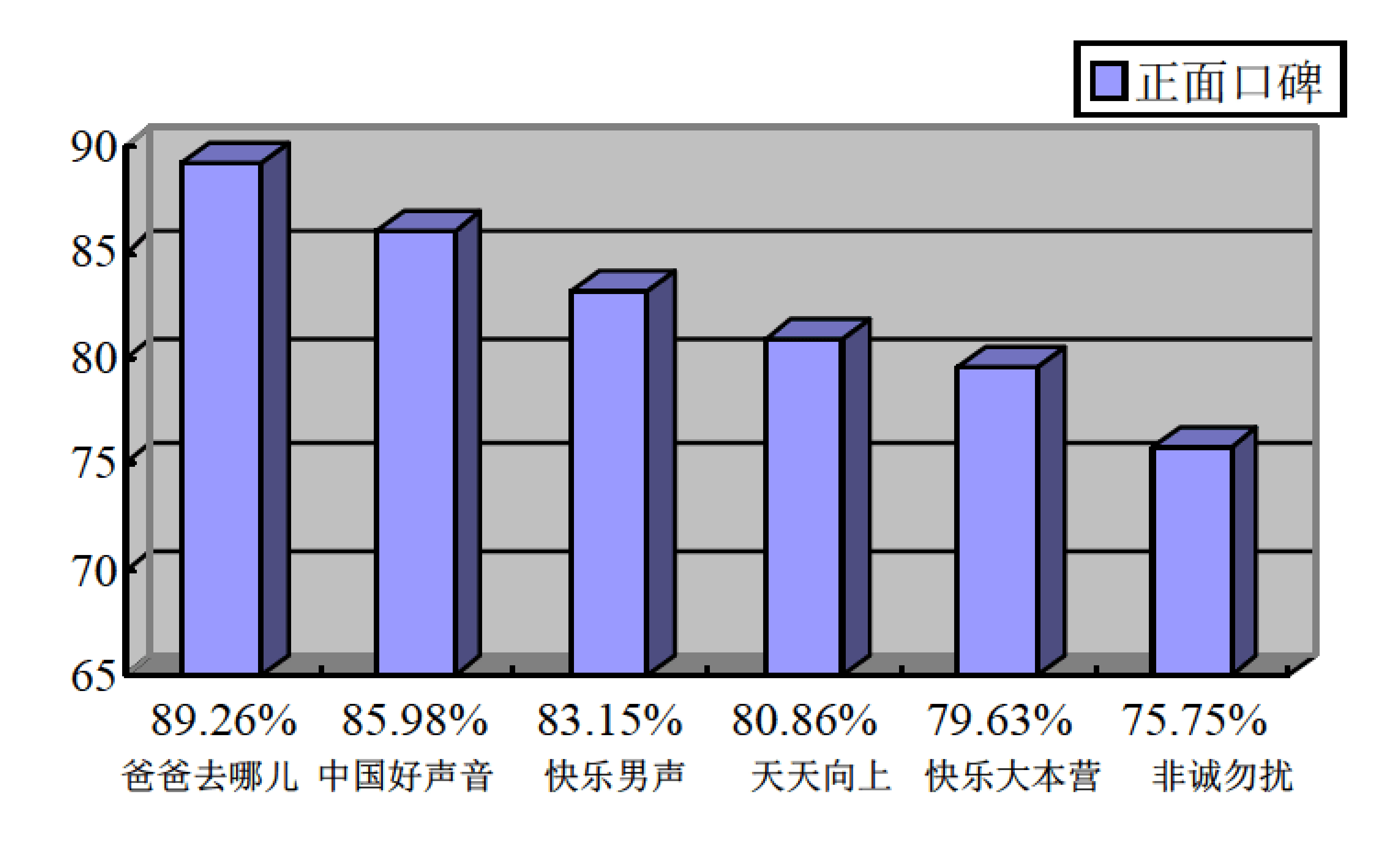


图 15-9 2013 年各热门电视节目口碑比较

湖南卫视《爸爸去哪儿》凭借“萌点”打动不少观众，几乎“零差评”的口碑令其收视较为突出，其中 CSM 全国网数据显示：收视率 1.1，市场份额 7.67%；CSM 29 城市网数据显示：收视率 1.46，市场份额 6.45%。在这两个收视数据网里，《爸爸去哪儿》均同时段第一。

【案例解析】：从本案例中可以看出，大数据的深入人心，或指明了未来电视必须从粗放式营销到精准营销转变的方向。对做内容产品来说，事先对数据掌握得越充分，未来在销售上就越有信心。例如，哪些人是你的忠实用户，哪些用户会根据节目产生消费行为，只有掌握这些数据，才能判断某种类型的节目适合做哪种产品。

由此可见，小作坊单打独斗的时代已经过去，只有坚持以数据为基础，掌握用户的喜好，再通过流程化的制作，才可能在互联网时代找到属于自己的立足之地。

15.3.2 【案例】用大数据来挖掘《小时代》

刚刚闭幕的第 16 届上海国际电影节又让“大数据”成为焦点，而郭敬明执导的电影《小时代》更是借助大数据的东风在上海国际电影节大出风头。

电影《小时代》讲述的是以经济飞速发展的上海为背景，4 个从高中就开始在一起生活的女生的故事。你可以讨厌《小时代》，但你却不能忽视《小时代》的观众群，因为他们或许将决定中国电影的未来。在一片争议声中，成本仅 2000 万元的《小时代》获得了接近 5 亿元的票房。按投资回报比计算，它甚至有望成为 2013 年“最赚钱”的华语电影。

数托邦工作室采用新媒体大数据分析手段，对《小时代》的观影人群进行了调查分

析。接下来就让我们从大数据的角度出发，“挖一挖”这部精确定位的所谓“脑残粉”电影的观影群体。数托邦工作室的数据采集方法如表 15-1 所示。

表 15-1　数托邦工作室的数据采集方法

步　骤	采集方法	具体数据
第一步	取样时间	2013-06-27 到 2013-07-01，即《小时代》上映之日起连续 5 天
第二步	抽样范围	每天抽取两万余条包含“小时代”关键词的微博，共采集到 106674 篇微博
第三步	用户抽样	从 106674 篇微博中抽取原发作者用户，去重后得到 100815 位用户
第四步	用户筛选	采用数托邦工作室的核心算法（准确率超过 90%），去除高度疑似“水军”账号 8670 个，去除机构账号 945 个，共保留 91200 位用户
第五步	群体微博	采集 9 万余位用户近期共约 900 万条有效微博

如图 15-10 所示，在《小时代》的 9 万多位微博原发作者中，女性占到了八成以上，接近半数还是微博达人，她们积极地参与了《小时代》这部电影的观影、评论、分享、传播甚至争论，创造了数倍于其他电影的有关《小时代》的各种微博。可见，她们既是《小时代》电影的主要观众群体，也对这部电影的传播和营销起到了至关重要的推手作用。

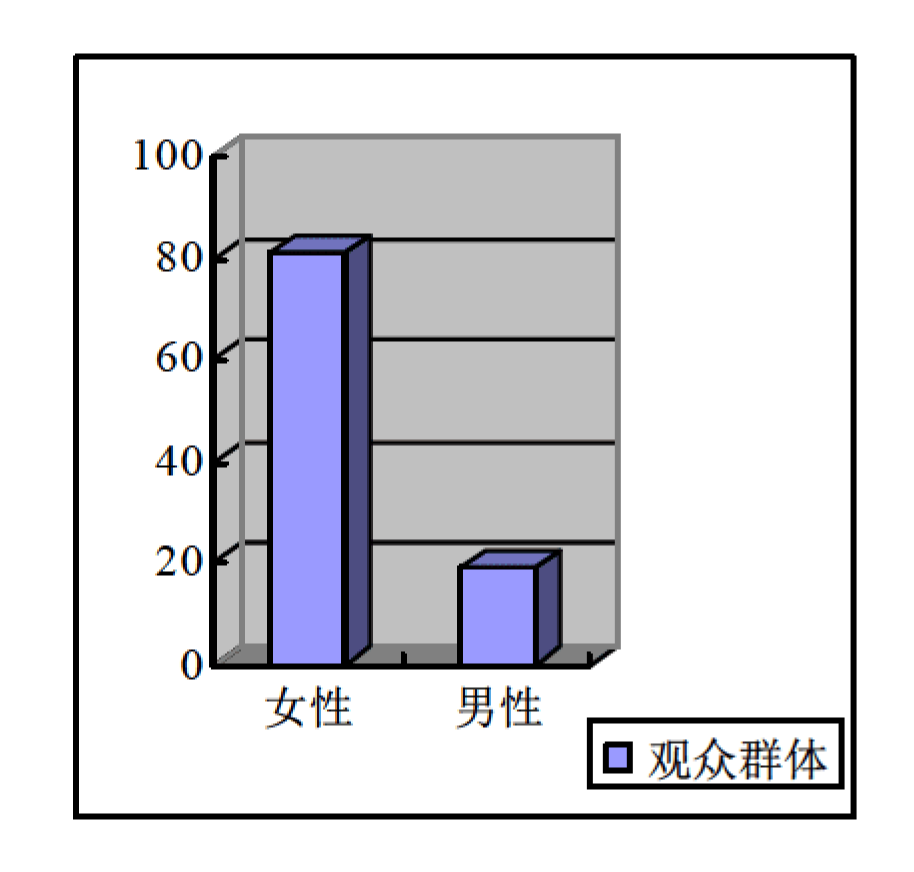


图 15-10　《小时代》的观众群体分析

【案例解析】：在本案例中，大数据分析扮演着一个针对影视制作及投资决策建议平台的角色，它可以提供对市场的理性预期，用精准的量化数字计算可能的投资回报率。大数据虽然解决不了艺术性的问题，但是却有商业借鉴意义。另外，大数据的分析还直接影响后期广告投放，以及衍生品的开发，有利于全价值链研究。

因此，笔者不得不承认，大数据对于当下电影创作起着至关重要的作用。尽管电影作为具有艺术属性的工业产品，无法用任何数据、技术手段取代，但除了创作之外的部分，如前期的观众导流、后期的宣传大多都是可以利用大数据去解决的。

15.3.3　【案例】《纸牌屋》变革传统电视业

大卫·芬奇的“导”和凯文·斯派西的“演”，无疑是美剧《纸牌屋》走红的关键原因。事实上，在两位重量级主创促成的成功背后，《纸牌屋》具有更多跨时代的意义——网站主导、数据先行。

出品方兼播放平台 Netflix 根据用户的数据总结收视习惯，并根据对用户喜好的精准分析来创作《纸牌屋》。《纸牌屋》的数据库包含了 3000 万用户的收视选择、400 万条评论、300 万次主题搜索。最终，拍什么、谁来拍、谁来演、怎么播，都由数千万观众的客观喜好统计决定。例如，在记录暂停、倒退、快进、评分、搜索的同时，进行大量截图，试图分析用户在音量、画面色彩甚至场景选取上的喜好。从受众洞察、受众定位、受众接触到受众转化，每一步都由精准细致高效经济的数据引导，从而实现大众创造的 C2B（Customer to Business，即消费者对企业），即由用户需求决定生产。

根据数据，点击率非常高的鬼才导演大卫・芬奇和男演员凯文・斯派西，成为了主创选择；再根据“政治惊悚”这类电影的受欢迎程度，Netflix 狠下心肠扔出了过亿美金，自制出了这部《纸牌屋》。

Netflix 将文艺创作一丝不苟地建立在对冰冷数据的分析上，而且达到了意想不到的好效果，《纸牌屋》迅速成为美国及其他 40 多个国家播出频率最高的电视节目，评论家毫不吝啬地给予它赞美之词，称之为“是一部艾美奖水准的电视剧”。

【案例解析】：在本案例中，《纸牌屋》的成功得益于 Netflix 海量的用户数据积累和分析。在任何一门生意中，能够预见未来都是可怕的，Netflix 在《纸牌屋》一战中可能已经接近这个水准。

如今，互联网以及社交媒体的发展让人们在网络上留下的数据越来越多，海量数据再通过多维度的信息重组使得企业都在谋求各平台间的内容、用户、广告投放的全面打通，以期通过用户关系链的融合，网络媒体的社会化重构，为广告用户带来更好、更精准的社会化营销效果。

笔者觉得，在不久的将来，大数据挖掘获得的结果也许比一个行业老手的直觉判断更准确。当然事情都有两面性，大数据分析在国内影视产业领域技术尚未成熟，但这恰恰是大数据在电影产业的机遇，也正是大量大数据分析技术人才的机遇，随着互联网的蓬勃发展以及中国电影产业的壮大，势必迎来大数据分析的春天。

专家提醒

当然，电影产业及市场还有很多影响因素，不仅仅是理性的数据分析，更有感性东西融入在电影中，但大数据对于电影产业的影响将会至关重要。

15.3.4 【案例】《纽约时报》让报纸智能化

《纽约时报》（The New York Times）作为一份享有世界声誉的报纸，是美国新闻界的领头羊和风向标。在 IT 技术的应用方面，《纽约时报》不惜重金打造智能商业系统，将围绕实时分析、智能预测和用户互动三大 IT 技术来提高新闻发布和时事分析的质量。

例如，位于加勒比海北部的海地发生大地震后，关于震情和救援的报道占据了各大

报纸和网站的首页。《纽约时报》将地震前后同一个地点的卫星地貌照片重叠放在了同一个窗口内，窗口内部有一个类似窗帘的分屏箭头，通过拉动它，读者可以看到同一个地点地震前后的变化。拉动分屏箭头的同时，还会自动浮现出相关的文字说明，如图 15-11 所示。



图 15-11 《纽约时报》关于海地大地震的报道页面

地震前，高尔夫球场一片翠绿；地震后，曾经翠绿的高尔夫球场，挤满了帐篷……通过这种对比，可以看到地表遭受的巨大破坏和当地灾民无家可归的惨状。和将两张地图简单地放到一起相比，这种信息表达方式增强了对比效果，使对比更加直观、一目了然。

【案例解析】：在本案例中，《纽约时报》通过对数据信息内容独具匠心的整合，把零散的信息融合为新的知识，产生了“1 + 1 > 2”的效果，给商务智能如何走向大众化提供了很好的启发。商务智能的应用注重信息的分析和整合，一个好的商务智能产品能够把复杂的信息内容视觉化、图像化、文字化，帮助用户看到不同事物之间的关系、联系以及发展的趋势和走向。

从《纽约时报》的案例可见，以构建 IT 运营平台为中心的时代即将过去，世界已经跨进了以数据分析和挖掘为中心的智能时代。

15.3.5 【案例】大数据带来逼真的影视特效

2012 年夏天上映的《百万巨鳄》是国内首部特效惊悚怪兽类型电影，片中的真正主角是一条名叫“阿毛”的长八米重达两吨的巨型鳄鱼。

巨鳄“阿毛”完全由特效制作产生，特效制作动物的关键就在于质感——皮肤的柔软度，牙齿、眼神等细小部位的刻画，稍有不慎就很容易露怯，如图 15-12 所示。为此，

制作方北京歌亮传媒有限公司召集了国内最顶级的特效技术人员，花了3个月的时间为鳄鱼形象作准备。特效制作过程分为多个工种，如建模、灯光、材质、渲染、动画、骨骼、肌肉动力学、特效、毛发等。其中，水和毛发的制作被认为是最难制作的特效种类，但这也是电影《百万巨鳄》中运用最多的部分。



图 15-12　利用大数据技术制作出逼真的巨鳄眼睛

《百万巨鳄》的拍摄和制作周期超过3年，其中大量的时间都花在了特效制作上。如何在有限的工期内高效地完成全片的特效制作工作，要求歌亮传媒的存储系统拥有更好的I/O处理能力和更高的数据吞吐量、更快的图片渲染和下载速度，大幅减少数据量大造成的系统处理瓶颈，从而实现更适合海量影像文件处理的数据管理、虚拟化和数据保护。

针对歌亮传媒的行业特点和应用需求，日立数据系统为歌亮传媒提供了适合于影视行业海量图片及非结构化数据信息处理的存储解决方案：以HNAS 3090为核心的数据处理解决方案，助力歌亮传媒实现对海量影像数据的高效管理以及基于底层的自动归档，从而有效提升了其IT系统能力，它不仅减少了后期制作人力消耗，更关键的是大大缩短了影片的上市时间。

通过大数据存储解决方案，歌亮的整个系统的数据读取速度得到了明显提升——可以同时为多人提供优越的读写服务，散文件读写也更加流畅，特效师和相关工作人员直接获得影像文件的速度提高了30%～40%，这大大提高了特效师们的创作效率，也不会让一些即兴的创作灵感因为数据调用的等待而消失殆尽。

【案例解析】：从《百万巨鳄》这部电影的实践来看，高精尖的数据专业技术人才对于电影的成功至关重要。电影的一个主要功能是娱乐大众，只有不断地制造出惊人和震撼的效果才能更好地实现电影的娱乐功能。在电影创作的过程中，技术无疑是最大的闪光点。

在商业社会中，“从数据中得到价值”一直都不是什么新鲜的东西，但是当大数据时代到来时，经济的新增量逐渐显露出来。如今，电影内容的创新与技术的创新已经融为一体。从电影产业的发展来看，像大数据这样的信息技术为电影创作提供支撑已经是大势所趋。

15.4 生活中的大数据应用案例

大数据，对普通老百姓而言，已经不再是一个陌生的词语。在这个海量信息的时代，大数据无时无刻不在影响、惠及、改变着我们的生活。我们在日常生活中所做的一切都会留下数字痕迹（或者数据），也就是大数据，我们可以利用和分析这些数据来让我们的生活更加美好。本节主要介绍大数据在生活方面的应用案例，希望对读者有一定的启发和学习价值。

15.4.1 【案例】大数据让你的生活更智能

我们经常会在匆忙外出的时候，忘记关闭正在使用的家电，例如电磁炉、空调等。回想起来的时候，心里不免惴惴不安，总是会犹豫是否要回家关闭。有强迫症的人们还会在外出时担心门是否已锁好等问题。

SmartThings 为我们提供了更智能的方法。SmartThings 公司可以帮助用户在家里安装动力、湿度和其他传感器，让你了解家里正在发生的事情，同时通过 iPhone 上的应用程序来控制家里的所有设备，如图 15-13 所示。



图 15-13 通过 iPhone 上的应用程序来控制家里的所有设备

SmartThings 采用了一种系统，可以将我们日常使用的实物连接到基于云的控制中心。该设备的重点指向是让一些终端设备连接至 SmartThings 中心，例如自动门锁、自动调温器、电源插座开关等。

例如，你可以用 SmartThings 来完成以下工作：

- 如果宠物跑出了院子，能够收到一个“哦，狗狗跑啦！”的通知。
- 如果浴室或者地下室发生了漏水事件，能够很快收到“漏水啦”的通知。
- 能够用“安全存储”功能，监控存放贵重物品的箱子或者抽屉是否被打开。
- 如果在社交网络里面收到新的粉丝或@时，能够通过手机上的渐变灯光提醒用户。

据悉，SmartThings 和现有的自动家用设备标准兼容，适用于数百个现有设备。

【案例解析】：在本案例中，SmartThings 通过收集家庭生活的种种数据，并利用云计算处理数据，可以使生活中的每样东西都变得智能。这样打开了无穷的可能性和无限的潜力，让用户的生活更加轻松、舒适和有趣。

15.4.2 【案例】数据能够开口说话当红娘

如果大数据能让谷歌预测出 2013 年美国将爆发流感，让微软公司成功预言奥斯卡金像奖 14 项大奖中的 13 项，那么大数据是否也能帮助单身者更快地找到对象呢？

2012 年底，网易旗下全新婚恋交友网站“花田”上线。“花田”以免费沟通为卖点，摒弃传统婚恋网站的“人工红娘”，从推荐到搜索全由系统自动完成。“花田”用大数据的精准化运营，为在海量异性资料中疲于搜索的用户“指一条明路”。

“花田”系统会自动推荐那些相对活跃、最近有信息流更新的人，这就促使用户拿出更新微博的劲头来更新“花田”，为其积累了大量可供分析的软性数据。目前，“花田”开发团队正试图通过自然语言处理技术和语义分析方法来解码用户性格，实现“软硬兼施”的精准推荐。

“花田”在对海量软硬数据进行分析的基础上，总结出一些人物特征，建立起一定数量的人物模型。再分析具体用户，将其分门别类套入各种模型。这样，用户心仪其中某一个人，便可向其推荐这一类人。“花田”试图将更高级的人脸识别，如五官识别、夫妻相匹配作为自己的增值服务收费点，对于此，尚有待进一步的技术突破。

“花田”还推出一个问答题库系统 Q&A，通过设置价值观、兴趣爱好、生活习惯、爱情观等分类问题，让用户参与答题。目前，花田平台预设 300 道 QA 题，已经有 20% 左右的用户拥有 Q&A 数据，平台用户答题数据量达到千万级。“花田”通过对 Q&A 数据的分析，能够发现两个异性之间在生活习惯、价值观、兴趣爱好等方面的契合度，建立数据模型，促使用户快速找到沟通的话题，如图 15-14 所示。



图 15-14　Q&A 数据的分析

自 2012 年 12 月 28 日向全国开放注册以来，“花田”注册用户已近 25 万，每日活跃用户达 4 万人。

【案例解析】：在本案例中，通过挖掘全站用户数据，并结合用户注册产品和使用产品的时间，网易“花田”可以精准地为用户推荐合适的匹配对象，就像是专门定制的一样。

专家提醒

数据分析不只可用于精准推荐，还能识别婚恋网站最为人诟病的造假和诈骗。例如，世纪佳缘的数据分析团队开发出一套网警系统，使自己由以往的被动等待用户举报骗子，改为主动出击。网警系统的原理是收集并分析骗子行为模式的数据，制作出一套骗子识别模型。

15.4.3 【案例】大数据保障人身财产安全

小说里的神探，不管是福尔摩斯、波洛，还是狄仁杰、柯南，都有一个共同的特点，那就是有一个具备强大分析能力的大脑，他们能够观察到细小的证据，并把这些证据关联起来，分析出犯罪事实。

目前，美国中央情报局已经开始利用大数据技术追踪恐怖分子和监控社会情绪。就像可口可乐等消费公司借助数据分析掌握消费者习惯一样，中情局也通过大数据技术来寻找恐怖分子的踪迹。此外，大数据分析可以了解多少人和哪些人正在从温和立场变得更为激进，并“算出”谁可能会采取对某些人有害的行动。

美国孟菲斯市警察局启用 Blue CRUSH 预测型分析系统后，使过去五年暴力犯罪率大幅下降。最近，美国马里兰州和宾夕法尼亚州也开始启用一种能极大降低凶杀犯罪率的犯罪预测软件，其不但能预测罪犯假释或者缓刑期间的犯罪可能性，还能成为法庭假释条款和审判的参考依据。

例如，美国加利福尼亚州圣克鲁兹市采用大数据算法可以计算出某时某地罪案（入室行窃、抢劫、偷车，但不包括杀人案）发生的几率。在过去两年中，该市的大约 100 名巡警在巡逻时会有针对性地出巡，他们携带的电子卡上会显示出附近最有可能发生罪案的 15 处地点。而在三分之二的情况下，大数据算法预测的罪案都确实发生了。

引入这个大数据算法后，圣克鲁兹市的入室行窃案件减少了 11%，偷车案减少了 8%，相应地，逮捕罪犯的成功率则提高了 56%。现在，美国已经有超过 10 市的警察局引入了这个大数据算法，其中包括洛杉矶、波士顿和芝加哥。

【案例解析】：在本案例中，大数据分析已经被用在刑事侦破领域，这为破获一些疑难杂案、保障老百姓的人身和财产安全提供了一种新的技术支持。其中，人脸识别技术的应用就是大数据挖掘的一个典型例子。

大数据分析的工具从长期来说，可以加速办案效率，优化警力资源分配，从而提高

社会和公众安全水平。随着警用大数据工具的不断成熟，以及“物联网＋社交网络＋大数据＋云计算”的高速融合发展，执法部门的犯罪侦破和预防将进入一个全新的大数据时代。

专家提醒

虽然大数据分析可以预测和阻止某些安全事故的发生，但事后的弥补也相当重要。大数据分析可被用来对过去事故评价分析，定位潜在的风险根源以及检测可导致安全事故的潜在苗头。

15.4.4 【案例】用大数据安全保管门钥匙

你是否遇到过不小心丢失或找不到钥匙的情况，如今找一位开锁匠来开门的话，除却高昂的人工费不说，还费时费力不安全。针对这一情况，纽约市有一家名为 KeyMe 的公司为大家带来了一个实用的解决方案——KeyMe 钥匙存储/复制机。

KeyMe 将该机器部署到了纽约市的 7～11 个便利店里面，有需要的人们可以选择“数字化”地复制并存储自己的钥匙，以便在紧急情况下迅速“还原”出一把备用钥匙。KeyMe 的外形类似于一台自动售货机，操作也非常简单，用户首先在线创建一个账户，然后机器会扫描钥匙并将其存储在云端，如图 15-15 所示。如果用户的钥匙不慎丢失，只需找到一台 KeyMe，通过指纹识别便可以迅速还原出一把钥匙，可选外形包括装饰性钥匙、组合型钥匙和开瓶器钥匙。



图 15-15　KeyMe

KeyMe 不会记录钥匙使用场景的信息，所有存储在云端的钥匙模型都只能通过指纹识别才能打开，而且创建 KeyMe 账户时还需要使用一张安全有效的信用卡。另外，每当有钥匙被还原出来时，系统都会自动给用户发一封验证邮件。

【案例解析】：在本案例中，KeyMe 的创意来自于大数据的云存储，其将每把钥匙的数据保存在云端。与 August 和 Lockitron 等智能锁相比，KeyMe 更加便携和兼容，

不需要电池，更不会崩溃。

对于使用云服务的企业来说，可以大大降低前期成本投入，并将更多的资金用在运营方面，而且由于不再需要自身去管理和维护服务器，他们会有更多的时间和精力专注于自身的主营业务。

15.4.5 【案例】地图 APP 成为生活好助手

笔者的好友李茂是个不折不扣的“地理白痴”，所以他下载了一个高德地图。只要花一点流量，李茂就能在地图上查看自己所处的位置，以及周围的建筑。

每天出门，你打开手机上的地图 App，运用实时交通功能，可以更顺畅地到达目的地。其实在简单的点点屏幕背后，有着看不见的复杂口令。在地图的实时交通领域，大数据令很多细碎繁琐的事情落地，“复杂”才决定了“简单”。

高德地图生产过程可分为三大环节：数据采集、数据生产、数据应用，如图 15-16 所示。高德在微信公众平台推出了服务号，可以供用户进行上下班路况查询，这同样是基于大数据的服务功能。

图 15-16　高德地图生产过程离不开大数据

笔者用一个实景比喻来解释大数据究竟能做到如何智能：当一个用户在某购物网站买过一张床后，他面对的不应该是隔三差五地收到同类产品的推荐信息，而是在几个月后收到特别为他定制的配套家具推荐。而在地图应用中，大数据同样要做到 2.0 版本，通过个性化的分析，做到量身定制的主动服务。例如，当你快下班时，就会收到一条推送信息，告诉你今天回家路上堵不堵，走哪条路最划算。

2013 年 5 月，阿里巴巴宣布对高德地图战略投资 2.94 亿美元，持有高德 28%的股份，成为高德第一大股东。阿里巴巴表示入股高德之后，会以移动互联网位置服务和深度生活服务的基础设施作为切入点，日后也将在数据建设、地图引擎、产品开发、云计算、推广和商业化等多个层面展开合作。

根据相关统计数据显示，高德导航地图在国内被广泛使用，占有 26%的市场份额。截止到 2013 年第一季末，这款应用每个月拥有 2900 万个活跃用户，而且总用户数在 5200 万以上。

【案例解析】：在本案例中，地图 APP 不仅能凭借大数据，为公众出行提供实时交通信息，还能整合生活服务，起到 O2O 总入口平台的作用。通过商家服务信息与地理信息的数据融合，地图将给用户带来更便捷的使用体验。

地图本身承载着各种各样的商业机构，无论是路边商店、实体店，还是日常生活中和吃喝玩乐、衣食住行相关的机构，都在地图上有所体现。因此，地图天然就具备承载各类生活服务的平台属性。

读书笔记

